REFERENCE ONLY

# UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author GRANT, A.D.

# Comparative Genome Analysis To Reveal Protein Evolution

Alastair Donald Grant

Biomolecular Structure and Modeling Unit
Department of Biochemistry and Molecular Biology
University College London

A thesis submitted to the University of London in the Faculty of Life Sciences for the degree of Doctor of Philosophy

**January 2006**





ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

# ABSTRACT

The completion of a substantial number of complete genome sequencing initiatives has produced more than a million protein sequences. Analysis of these protein sequences is possible using recent advances in computing and bioinformatics techniques.

This thesis describes a novel automated protein classification protocol which groups proteins into families and identifies protein domain architectures via domain assignment. This data is presented in the Gene3D database which is used for subsequent analysis.

The analysis of the distribution of protein family and protein domain data shows a power-law like distribution that is typically seen in many biological data distributions and is indicative of the small world networks that underlie biological systems biology. Kingdom distribution of superfamilies and protein families in Gene3D has been used to describe the evolutionary mechanisms that determine genome diversity through protein diversity. Domain occurrence profiles have been used to identify protein domain superfamilies that are correlated with genome size in bacteria. These superfamilies are shown to exhibit a balance between metabolic and regulatory roles along microeconomic principles that may determine bacterial genome size.

Domain families identified in Gene3D enable a determination of the total number of protein folds in nature. Sub-clustering of domain families permits domain family sub-cluster occurrence profiles to be determined. These profiles are shown to be capable of detecting correlations and anti-correlations between domain families that are undetectable using superfamily occurrence profiles alone. Clusters of correlated domain subclusters are shown to identify functionally linked clusters of proteins. Finally, the data in Gene3D is used to functionally annotate the CATH database and provide functional predictions for un-annotated proteins, providing more comprehensive functional repertoire and greater accuracy than other functional prediction methods.

# ACKNOWLEDGEMENTS

I would like to thank all my colleagues, especially those at CATH, including: Sarah Addou, Adrian Akpor, Chris Bennett, James Bray, Dan Buchan, Tim Dallman, Ilhem Diboun, Lesley Greene, Andrew Harrison, Caroline Johnston, Nick Keep, Tony Lewis, Stefano Lise, Francis Pearl, Stathis Sideris, Antonio Sillero, Ian Sillitoe, Annabel Todd, Ollie Redfern, Gabby Reeves, and Corin Yeats. Working with all these people has been really interesting and I have enjoyed it immensely. Particular mention must go to Dave Lee, for introducing bioinformatics of genomic proportions; Juan Ranea, for complicating everything even further, Mark Dibley, for keeping everything together behind the scenes; and finally Michael Maibaum and Russell Marsden for many helpful discussions.

Special thanks must go to the system administrators, without whom all things would promptly grind to a halt: Jahid Ahmed, Donovan Bins, Jesse Oldershaw, and especially Duncan McKenzie, for repeatedly reviving so many dead machines.

This thesis would not have been at all possible without the dedication, vision and support of my supervisor, Christine Orengo. I will never be able to thank her enough, for her insight and guidance, and for providing me with the opportunity to work in such a fantastic laboratory. I would also like to thank my mentor, Paul Driscol for his support, and the MRC for funding my studentship.

Thanks to my parents Chris and Kathy, and my siblings Colette and Hugh Grant, for their love. My friends, of whom only Lillian Tsang, Abigail Jones, Karine Rousseau, Ashvin Mathoora, Robin Mukerji, David 'Wavey' Holroyd and Sena Gbeckor-Kove get to see their names in print, and all the folks in BoA.

Finally, my gorgeous wife Ronit, I love you! A special mention goes to my son, Amir, whose hugs and kisses, and occasional words, have provided many a happy diversion in the past twenty months.

*For Ronit, As You Wish*

# CONTENTS

**ABSTRACT** ..... 2

**ACKNOWLEDGEMENTS** ..... 3

**List of Figures** ..... 9

**List of Tables** ..... 11

**CHAPTER ONE** ..... 12
**Introduction** ..... 12
**1.1 Introduction** ..... 12
**1.1.1 Deoxyribonucleic Acid** ..... 12
**1.1.2 Transcription and Translation** ..... 15
**1.1.3 Gene Identification** ..... 16
**1.1.4 DNA Sequencing** ..... 18
**1.1.5 Genome Sequencing** ..... 19
**1.1.6 Sequence Databases** ..... 19
**1.1.7 Protein Structure Determination** ..... 20
**1.1.8 Protein Structure** ..... 21
**1.1.9 Structure is more Conserved than Sequence** ..... 23
**1.1.10 Protein Structure Classification** ..... 26
**1.2 Homology** ..... 29
**1.2.1 Sequence Based Homology Detection Methods** ..... 29
*1.2.1.1 BLAST1* ..... 30
*1.2.1.2 BLAST2* ..... 30
*1.2.1.3 Expectation Values* ..... 31
*1.2.1.4 Hidden Markov Models* ..... 32
*1.2.1.5 Comparison of BLAST1, BLAST2 and HMMs* ..... 33
**1.2.2 Structure Based Homology Detection Methods** ..... 34
*1.2.2.1 Structural Alignment* ..... 35
*1.2.2.2 Threading* ..... 35
**1.2.3 Context Based Functional Prediction Methods** ..... 36
*1.2.3.1 Rosetta Stone* ..... 36
*1.2.3.2 Protein-Protein Interaction* ..... 37
*1.2.3.3 Synteny* ..... 37
*1.2.3.4 Phylogenetic Profiles* ..... 38
*1.2.3.5 Expression Profiles* ..... 38
**1.3 Functional Annotation** ..... 39
**1.3.1 Defining Function** ..... 39
**1.3.2 Enzyme Commission** ..... 40
**1.3.3 Kyoto Encyclopaedia of Genes and Genomes** ..... 40
**1.3.4 Clusters of Orthologous Groups** ..... 40
**1.3.5 Genome Ontology** ..... 41
**1.3.6 Affymetrix** ..... 41
**1.3.7 STRING** ..... 41
**1.3.8 Reliability of Annotation** ..... 42
**1.4 Objectives** ..... 43

**CHAPTER TWO** ..... **44**
**Construction of the Gene3D Resource of Complete Genomes Annotated with Protein Family, Domain Family and Functional Information** ..... **44**
**2.1 Introduction** ..... **44**
**2.1.1 The Repertoire of Completed Genomes** ..... **44**
**2.1.2 Protein Annotation** ..... **45**
**2.1.3 Protein Clustering Methods** ..... **45**
*2.1.3.1 Subclustering Families* ..... *46*
**2.1.4 Homology Detection** ..... **47**
*2.1.4.1 Building Hidden Markov Models using SAMT* ..... *48*
**2.1.5 Protein Family Resources** ..... **49**
*2.1.5.1 Sequence Based Protein Family Resources* ..... *50*
*2.1.5.2 Families of Whole Protein Sequences* ..... *52*
*2.1.5.3 Families of Protein Domain Sequences* ..... *54*
*2.1.5.4 Structure Based Protein Family Resources* ..... *55*
**2.1.6 Structural Annotation of Genomes** ..... **58**
**2.2 Objectives** ..... **60**
**2.3 Results** ..... **61**
**2.3.1 Genome Sources in Gene3D** ..... **61**
**2.3.2 Domain Sources in Gene3D** ..... **61**
**2.3.3 Family Clusters in Gene3D** ..... **62**
**2.3.4 Database Tables in Gene3D** ..... **62**
**2.4 Protein Family Landscape Protocol** ..... **64**
**2.4.1 Stage 1: Protein Family Clustering** ..... **65**
*2.4.1.1 Benchmarking TribeMCL using Structural Data* ..... *66*
**2.4.2 Stage 2: Domain Assignment** ..... **68**
*2.4.2.1 Building HMM Libraries* ..... *69*
*2.4.2.2 Benchmarking HMM Libraries* ..... *70*
*2.4.2.3 Domain Assignment by DomainFinderII* ..... *71*
*2.4.2.4 Percent Model Matched Domain Assignment Threshold* ..... *72*
*2.4.2.5 Acceptable Overlap Domain Assignment Threshold* ..... *73*
*2.4.2.6 E-value Domain Assignment Cut-Off* ..... *75*
*2.4.2.7 Resolving Multiple Overlapping Assignments* ..... *75*
*2.4.2.8 Domain Architectures* ..... *77*
**2.4.3 Stage 3: Functional Annotation** ..... **78**
*2.4.3.1 Functional Assignment in Gene3D* ..... *78*
*2.4.3.2 Function Assignment to Gene3D Proteins* ..... *78*
*2.4.3.3 Functional Coverage of Genomes* ..... *80*
**2.5 User Interface** ..... **82**
**2.6 Summary** ..... **85**

**CHAPTER THREE** ..... **86**
**Analysis of Protein Families and Domain Families in 120 Complete Genomes** ..... **86**
**3.1 Introduction** ..... **86**
**3.1.1 Power Laws in Protein Family Data** ..... **86**
**3.1.2 Novel Protein Families** ..... **87**
**3.1.3 Domain Assignment to Genomes** ..... **88**
*3.1.3.1 Un-assignable Regions* ..... *88*
**3.1.4 Domain Architecture** ..... **89**
**3.2 Objectives** ..... **90**
**3.3 Results** ..... **91**
**3.3.1 Analysis of Protein Family Populations in Gene3D** ..... **91**

*3.3.1.1 Size Distribution of Protein Families* ..... *91*
*3.3.1.2 Diversity of Protein Families in Gene3D* ..... *93*
*3.3.1.2.1 Function of Invariant Protein Families* ..... *94*
*3.3.1.2.2 Function of Diverse Protein Families* ..... *95*
**3.3.2 Analysis of Domain Family Populations in Gene3D** ..... **97**
*3.3.2.1 Size Distribution of Domain Families* ..... *97*
**3.3.3 Domain Assignments to Protein Families in Gene3D** ..... **101**
*3.3.3.1 CATH and Pfam Domain Assignment Overlap* ..... *103*
**3.3.4 Domain Architectures in Gene3D** ..... **104**
*3.3.4.1 Domain Architecture Consistency in Protein Families* ..... *105*
*3.3.4.2 Domain Architecture Superfamilies in Gene3D* ..... *106*
**3.3.5 Using Gene3D Families to Validate Genscan Predictions** ..... **107**
**3.3.6 Genome Coverage in Gene3D** ..... **108**
**3.3.7 Increasing Genome Coverage in Gene3D** ..... **112**
*3.3.7.1 HMM Library Expansion* ..... *112*
*3.3.7.2 Updated Versions of CATH and Pfam* ..... *113*
**3.3.8 Kingdom Distribution of Protein Families and Domain Families in Gene3D** ..... **115**
**3.4 Summary** ..... **118**

**CHAPTER FOUR** ..... **119**
**Application of Gene3D to Structural Genomics** ..... **119**
**4.1 Introduction** ..... **119**
**4.1.1 How many Domain Families are Currently Recognised and how many Novel Folds can we predict using this data?** ..... **122**
**4.2 Objectives** ..... **126**
**4.3 Results** ..... **127**
**4.3.1 Calculating the Number of Domain Families in Gene3D** ..... **127**
**4.3.2 Should Structural Genomics be Targeting Singletons?** ..... **129**
**4.3.3 How many folds remain to be discovered by Structural Genomics?** **129**
**4.3.4 Structural Genomics Target Selection Using Gene3D** ..... **130**
*4.3.4.1 Coarse Grained Target Selection to Identify Novel Folds* ..... *130*
*4.3.4.2 Fine Grained Target Selection to Increase the Number of Homology Models for Genome Sequences* ..... *134*
*4.3.5 Prioritising Sequence Diverse Domain Families* ..... *135*
*4.3.6 Prioritising Functionally Diverse of Domain Families* ..... *139*
**4.4 Summary** ..... **144**

**CHAPTER FIVE** ..... **145**
**Phylogenetic Occurrence Profiles to Analyse the Function and Evolution of Domain Families** ..... **145**
**5.1 Introduction** ..... **145**
**5.2 Objectives** ..... **147**
**5.3 Results** ..... **148**
**5.3.1 Analysis of the Genome Size Dependence of CATH Superfamilies** **148**
**5.3.2 Identification of Universal CATH Homologous Superfamilies** ..... **150**
**5.3.3 Distribution of Size-Dependent Universal CATH Superfamilies** ..... **151**
**5.3.4 Analysis of the Function of Size-Dependent Universal CATH Superfamilies** ..... **153**
**5.3.5 Identifying the Bacterial Genome Size Determinants of Size Dependent Universal Superfamilies** ..... **154**
*5.3.5.1 Economies of Scale* ..... *154*
*5.3.5.2 Predicting Optimal Bacterial Genome Size* ..... *155*

**5.4 Using Gene3D Phylogenetic Occurrence Profiles for Predicting Protein Functional Relationships ..... 159**
**5.4.1 Pair Comparison of Profiles ..... 160**
**5.4.2 Degenerate Domain Family Subcluster Profiles ..... 162**
**5.4.3 Information Content of Profiles ..... 162**
**5.4.4 Comparison of Gene3D Profiles to Randomised Null Models ..... 163**
**5.4.5 Gene3D Phylogenetic Occurrence Profile Clustering ..... 164**
**5.4.6 Functional Clusters revealed by Gene3D Phylogenetic Occurrence Profile Clustering ..... 165**
*5.4.6.1 Profile Clusters Representing known Functional Groups ..... 165*
*5.4.6.2 Deep Domain Family Subcluster Profiles ..... 168*
**5.4.7 User Defined Query Profiles ..... 169**
**5.5 Summary ..... 172**

**CHAPTER SIX ..... 173**
**Discussion and Future Work ..... 173**
**6.1 Discussion ..... 173**
**6.2 Future Work ..... 176**

**APPENDICES ..... 178**

**REFERENCES ..... 189**

## List of Figures


1.0 Structure of DNA

1.1 Central Dogma

1.2 Gene Structure

1.3 Protein Structure

1.4 Sequence Identity versus Structure Similarity

1.5 Structure is more Conserved than Sequence

1.6 Functional Conservation versus Sequence Identity in CATH

1.7 Protein Structure Classification

1.8 Linear HMM

1.9 Structural Similarity in the Absence of Sequence Similarity

2.0 Increase in Genomic Data

2.1 Problems Associated with Clustering

2.2 Iterative Nature of the HMM Build Process

2.3 Gene3D Database Table Structure

2.4 PFscape Protocol Structure

2.5 TribeMCL Granularity Benchmarking

2.6 TribeMCL Granularity Benchmarking

2.7 HMM Representation of a CATH Homologous Superfamily

2.8 HMM Coverage of CATH Homologous Superfamilies

2.9 Percent Model Matched Cut-off

2.10 Acceptable Overlap in DomainFinderII Domain Assignment

2.11 DomainFinderII Effect

2.12 Domain Architecture Assignment Protocol

2.13 Gene3D Coverage of Affymetrix Microarray

2.14 Functional Annotation of Proteins in Gene3D

2.15 Functional Annotation of Genomes in Gene3D

2.16 Gene3D Website

2.17 Gene3D Domain Assignment Diagram

3.0 Power Law Distribution of Protein Families

3.1 Frequency Distribution of Protein Family Diversity

3.2 Functions of Invariant Protein Families in Gene3D

3.3 Functions of Diverse Protein Families in Gene3D

3.4 Sizes of Domain Families

3.5 Log-log Plots of (a) CATH, (b) Pfam and (c) Newfam Families

3.6 Length Distribution of Newfam Sequences compared to CATH Domains
3.7 CATH Domain Coverage of Protein Families
3.8 Pfam Domain Coverage of Protein Families
3.9 Overlap of CATH and Pfam Domain Assignments
3.10 Consistency of Domain Architecture in Protein Families
3.11 Domain Architecture Distribution across Protein Families
3.12 Gene Coverage of Genomes in Gene3D
3.13 Residue Coverage of Genomes in Gene3D
3.14 Kingdom Distribution of Domains in Gene3D
3.15 Kingdom Distribution of Protein Families in Gene3D
4.0 The Number of Targets at each stage of the Structural Genomics Pipeline
4.1 Percentage of CATH Folds Accounting for Percentage of CATH s35 Sequence Families in Gene3D
4.2 Running Total of the Percentage of Pfam Domain Sequences in the largest Pfam families in the Genomes
4.3 Running Total of the Percentage of Domain Sequences in CATH, Pfam and Newfam families in the Genomes
4.4 Size/Diversity of CATH and Gene3D Domain Families
4.5 Relative Diversity of CATH Domain Families
4.6 Functional Characterisation of Domain Families
4.7 Functional Diversity versus Sequence Diversity of CATH Domain Families in Gene3D
5.0 Domain Occurrence Profiles
5.1 Size-Dependent CATH Homologous Superfamilies
5.2 Universal Size-Dependent Superfamilies
5.3 Distribution of Size-Dependent Universal Superfamilies
5.4 Economies of Scale: Optimum Factory Size
5.5 Economies of Scale: Optimum Bacterial Genome Size
5.6 Distribution of Bacterial Genome Size
5.7 Building Gene3D Phylogenetic Profiles
5.8 Profile Pair Comparison
5.9 Comparison of Gene3D Profile Correlations to Randomised Null Model Datasets
5.10 Gene3D Profile Clustering
5.11 Eukaryotic Cluster Domain Occurrence Profiles

## List of Tables


2.0 Summary of Example Sequence Family Databases

2.1 Protein Structure Family Resources

2.2 Functional Annotation by Kingdom in Gene3D

2.3 User Interface Queries in Gene3D

3.0 Number of Protein Families and Subclusters in Gene3D

3.1 Largest Protein Families in Gene3D

3.2 Largest Domain Families in Gene3D

3.3 Average Domain Coverage in Gene3D

3.4 Genome Coverage in Gene3D

3.5 Escherichia coli Genome Coverage using s35 and s95 HMM Libraries

3.6 Genome Coverage in Gene3D using various HMM Libraries

4.0 Domain Family Characterisation in Gene3D

4.1 Most Diverse CATH Domain Families in Gene3D

4.2 Most Structurally Under-represented CATH Domain Families

4.3 Scope of Functional Annotation in Gene3D Families

4.4 Ten Structurally and Functionally Diverse CATH Domain Families Dominate Gene3D

5.0 Depth of Profiles in Eukaryote/Prokaryote Dataset

5.1 Profiles in the Ras Query Cluster

# CHAPTER ONE

# Introduction

## 1.1 Introduction

Biology has entered a genomic age. The development of high-throughput automated experimental techniques has uncovered biological data at a faster rate than ever before and this looks likely to continue for the foreseeable future. The huge amount of data generated daily necessitates the use of computational methods for organisation, classification and analysis. Biology is exploring complex biological networks by genome level investigation. Functional and comparative genomics aim to understand how species have evolved and determine the function of proteins and non-coding genomic regions, primarily through identifying and comparing homologues. Structural genomics aims to use the results from genome sequencing projects and advances in structural determination to define fold space through organisation and analysis of protein structures. Resources which classify protein structures and family relationships across multiple genomes can provide useful information for functional, comparative and structural genomics. This chapter introduces the biological molecules which are core to evolution and function in organisms, and methods that have been developed to compare and describe their different physical characteristics, genomic context and biological functions.

### 1.1.1 Deoxyribonucleic Acid

In 1944, Avery, MacLeod and McCarty (Avery *et al.*, 1944) showed that the molecule deoxyribonucleic acid (DNA) carried inheritable information by building on the work of Frederick Griffith who had showed in 1928 that heat-killed virulent bacteria could transfer their virulence to non-virulent bacteria.

The observation of the relative levels of nucleotide bases in DNA, made by Erwin Chargaff in 1950, showing that the amount of adenine was equal to the amount of thymine, and that the amount of guanine was equal to the amount of cytosine, combined

with the insight by Rosalind Franklin (Franklin and Gosling, 1953) and Maurice Wilkins (Wilkins *et al.*, 1953) that the DNA molecule was shaped like a helix, containing two 'strands' joined by 'rungs', led to the discovery of the molecular structure of DNA in 1953 by Francis Crick and James Watson (Watson and Crick, 1953).

The DNA molecule consists of two polymeric strands in a right-handed helix, joined together by hydrogen bonds between complementary nucleotide base pairs. Each polymer strand consists of nucleotides, which are made of a phosphate group bound to a deoxyribose sugar, which is bound to a nitrogenous base. There are four different kinds of nitrogenous base in DNA; these are the pyrimidines, cytosine (C) and thymine (T), and the larger purines, adenine (A) and guanine (G). Polymerisation of nucleotide subunits occurs with covalent phosphodiester bonds formed between the 5' and 3' hydroxyl groups on the deoxyribose sugar and phosphate groups, forming an alternating sugar-phosphate backbone. Base pairs are formed between a complementary purine and a complementary pyrimidine, where adenine always binds to thymine and cytosine always binds to guanine.

Two anti-parallel polynucleotide chains form a DNA molecule with hydrophilic sugar-phosphate backbones on the outside of the helix and the hydrophobic hydrogen bonded base pairs stacked perpendicular to the helix axis, on the inside of the helix (shown in figure 1.0). The DNA helix can adopt three different conformations in nature, depending on base pair composition and hydration/ion levels. The common form is the B-helix, with ten nucleotides per helix turn and both a major and minor groove present on the helix surface. G-C rich DNA forms Z-helices, where the helix becomes left-handed, longer, and thinner, with twelve nucleotide bases per helix turn and only a single groove on the helix surface. Lastly the A-helix conformation is seen at low hydration levels or high cation concentrations. This conformation has eleven nucleotides per helix turn and two grooves present on the helix surface.

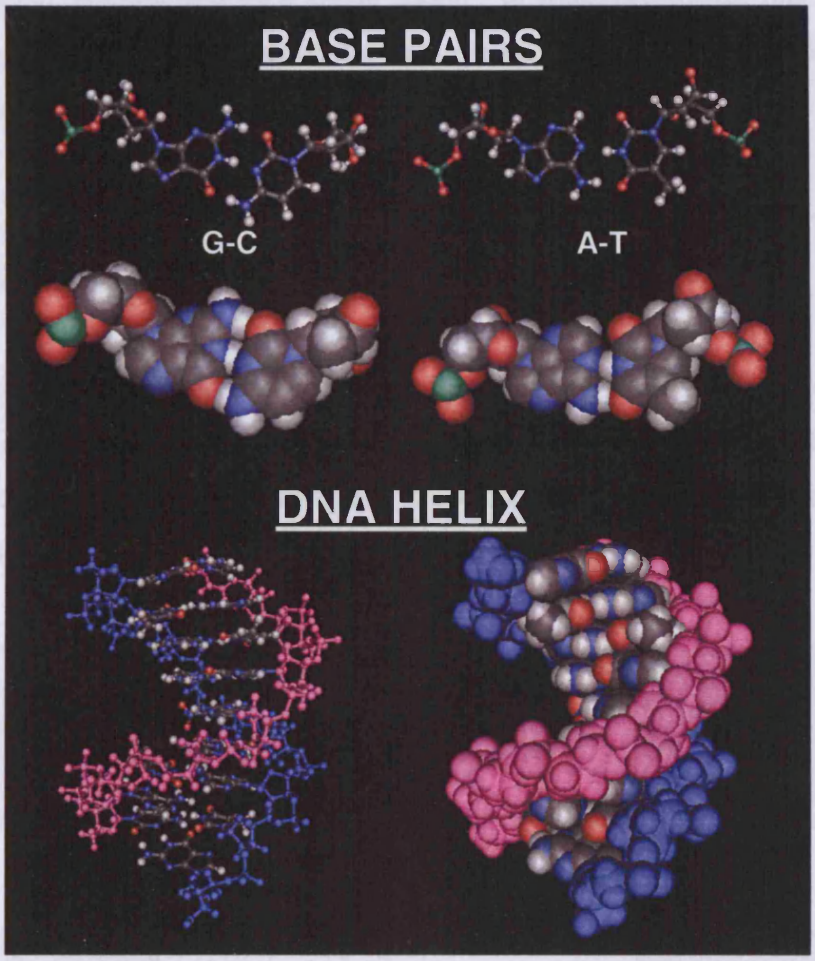


**Figure 1.0 Structure of DNA.** *Ball-and-stick and space-filled models showing the structure of nucleotide base pairs cytosine-guanine (G-C, top left), adenine-thymine (A-T, top right) and the double-helical structure of the deoxyribonucleic acid molecule (bottom). Figure shows structure of PDB 1DK6 (Klewer et al., 2000).*

The structure of DNA, consisting of complementary sequences of the four nucleotide base pairs led George Gamov (Gamov *et al.*, 1956) to postulate that to encode the twenty essential amino acids would require at least nucleotide base triplets to encode each amino acid. Experiments by Nirenberg and Matthaei in 1961 (Nirenberg and Matthaei, 1961), and later by Nirenberg and Leder in 1964 (Nirenberg and Leder, 1964; Leder and Nirenberg, 1964), showed that the genetic code consists of sixty-one codons (nucleotide triplets), encoding specific amino acids and three codons encoding termination (of the process of translation). The genetic code has built in redundancy, in that multiple codons can encode some amino acids. Only two amino acids, methionine

and tryptophan, are encoded by a single codon. In some organisms, one or two stop codons may also encode amino acids, for example in many genomes, including humans, one of the stop codons sometimes encodes selenocysteine.

### 1.1.2 Transcription and Translation

There are many differences in prokaryotic and eukaryotic nuclear material packaging and cellular structure, but both prokaryotes and eukaryotes employ the same genetic code and translation/transcription mechanisms, known as the Central Dogma (shown in figure 1.1), to express protein from DNA:

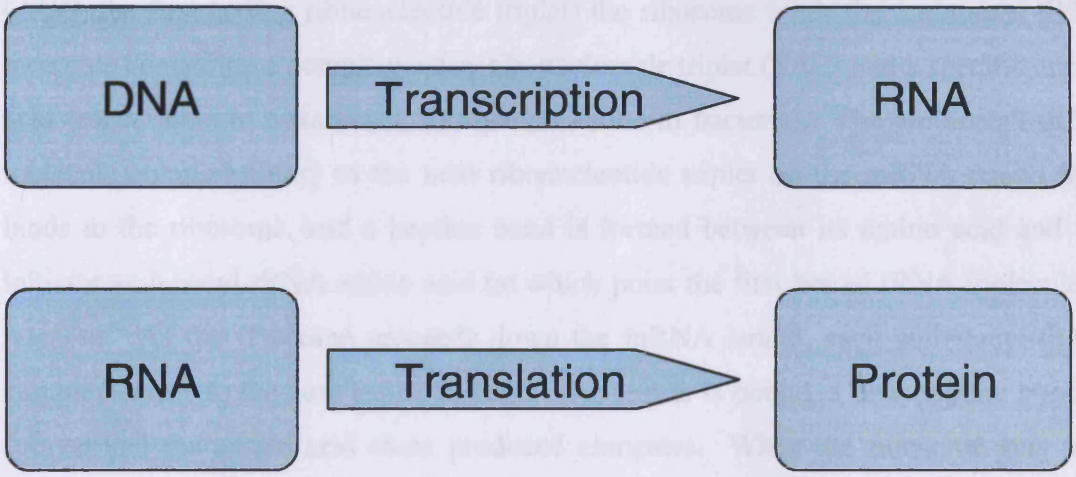


**Figure 1.1 Central Dogma.** *In cellular systems, genetic information transfers from DNA to RNA to Protein, and cannot be transferred from Protein to either Protein or DNA or RNA*

The triplet codes in DNA are transcribed into RNA by RNA polymerase. This enzyme binds to transcription factors that bind to promoter regions 5' terminal to coding DNA, which open the helix to allow the RNA polymerase to proceed in a 3'-5' direction down one polynucleotide strand of the DNA molecule. In eukaryotes, where DNA is packaged into protein complexes, these complexes are unwound and then re-wound once the RNA polymerase has passed. As the RNA polymerase proceeds down the DNA strand it matches complimentary ribonucleotides to the sequence of nucleotides in the DNA strand. Polymerisation of these ribonucleotides forms a 5'-3'

ribonucleic acid polymer (RNA) strand, complimentary to the 3'-5' DNA strand. There are several types of ribonucleic acid: mRNA functions as a coding template for translation; rRNA functions as part of the ribosome, an rRNA/protein complex required for translation; and tRNA which bind amino acids to form aminoacyl-tRNA, used for translation, which possess specific ribonucleotide triplet codes linked to a specific amino acid.

In prokaryotes and about 10-15% of translation in eukaryotes, translation occurs simultaneously with transcription. The remaining translation in eukaryotes occurs outside the nucleus, and so mRNA is first transported to the cytoplasm. Translation begins when the ribosome binds 5' terminal to the coding region of the mRNA strand. The ribosome proceeds in a 5'-3' direction down the mRNA strand, at the start codon (AUG, the first coding ribonucleotide triplet) the ribosome binds the aminoacyl-tRNA molecule containing a complementary ribonucleotide triplet (TAC) and a specific amino acid (methionine in eukaryotes, formylmethionine in bacteria). The aminoacyl-tRNA molecule complementary to the next ribonucleotide triplet on the mRNA strand then binds to the ribosome, and a peptide bond is formed between its amino acid and the initiator aminoacyl-tRNA amino acid (at which point the first bound tRNA molecule is released. As the ribosome proceeds down the mRNA strand, each aminoacyl-tRNA complementary to the next triplet in the coding region is bound, a new peptide bond is formed and the amino acid chain produced elongates. When the ribosome gets to a termination triplet, no aminoacyl-tRNA molecule is bound and both the ribosome and the newly translated peptide are released.

### 1.1.3 Gene Identification

The region of DNA encoding an RNA molecule is called a gene. The peptide sequence, the primary structure of a protein, can be identified by translating the genetic code from a DNA sequence encoding a gene (shown in figure 1.2). The coding region in prokaryotic and several eukaryotic genes is a continuous open reading frame of triplets, but the vast majority of eukaryotic genes have discontinuous coding regions, where coding exons are separated by non-coding introns.

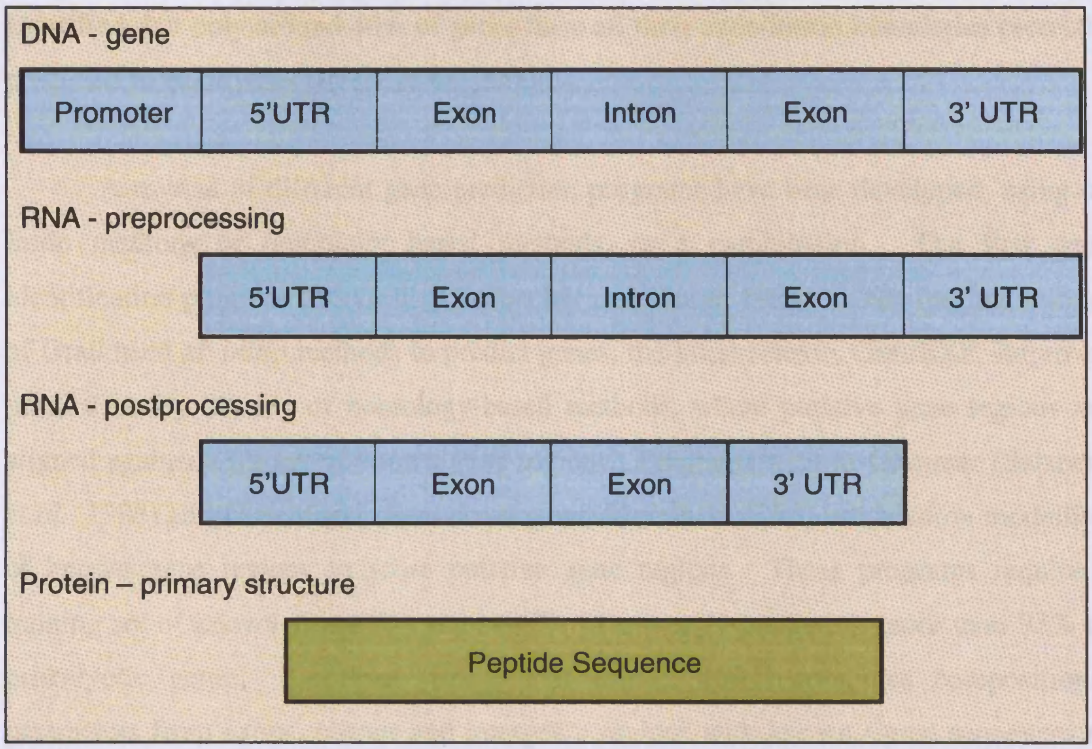


**Figure 1.2 Gene Structure.** *DNA sequence regions are transcribed into RNA, which is processed to remove non-coding intronic sequences. Translation of the resulting RNA produces the protein's primary structure.*

In prokaryotes the continuous coding regions make peptide sequence prediction from DNA sequences relatively straightforward. Identification of start codons and stop codons, and translation of the intervening coding region allows the vast majority of prokaryotic proteins to be predicted. To reduce erroneous protein predictions caused by DNA sequencing errors and shadow reading frames, additional evidence that the DNA sequence actually does encode a protein is desirable. Sequence similarity or similar sequence characteristics (codon bias, G-C content) to known expressed proteins or identification of a ribosomal binding site or promoter region increase the reliability of protein predictions.

The presence of exons and introns in eukaryotic genes, which often produces complex and diverse mRNA variations, makes accurate peptide sequence prediction more of a challenge. In some eukaryotes, for example the yeast *Saccharomyces cerevisiae*, only about 5% of genes contain introns (Patthy, 1999) whilst in other eukaryotes intronic genes are far more widespread. Up to 95% of coding DNA can be

identified, but only around 40% of genes have all their exon/intron boundaries correctly predicted in eukaryotes (Reese *et al.*, 2000).

A myriad of different gene prediction programs have been developed, using *ab initio* methods or homology based methods, or a combination. The first gene identification program was Grail (Uberbacher and Mural, 1991). The original release of Grail used *ab initio* methods to predict genes, the latest release, GrailEXP, improves predictions by the use of homology-based methods, where putative gene regions are aligned against a library of known gene regions. Programs such as Glimmer (Salzberg *et al.*, 1998) and GeneMarks (Borodovsky and McIninch, 1993) use Markov modelling of known gene regions to score putative gene regions. These programs require a training set of known genes, but are capable of correctly predicting more than 98% of prokaryotic genes. GenScan (Burge and Karlin, 1997) combines compositional parameters from exons, introns and intergenic regions with known signal sequences to predict genes.

### 1.1.4 DNA Sequencing

Two methods of DNA sequencing, Maxam-Gilbert (Maxam and Gilbert, 1977) and Sanger-Coulson (Sanger *et al.*, 1977) were independently developed in 1977. Both methods used four independent reactions to identify the four different types of nucleotide bases in DNA. Maxam-Gilbert sequencing labels DNA with 32-P; nucleotide base-specific chemical degradation then produces DNA fragments of varying length. Sanger-Coulson sequencing labels DNA with 35-S; enzymatic synthesis using nucleotide base-specific didesoxylribonucleotide terminators produces DNA fragments of varying length.

These DNA fragments are gel-electrophoresised in four lanes, each lane corresponding to chemical degradation or termination reaction for each nucleotide base in DNA. A radiographic film of the gel can then be used to determine the sequence of nucleotide bases in the DNA. In 1987, Prober *et al.* (Prober *et al.*, 1987) modified the Sanger-Coulson sequencing method using fluorescent dideoxyribonucleotide terminators. This enabled a single reaction to identify all four nucleotide bases, since each could be labelled a different colour, and permitted single lane gel-electrophoresis.

In addition this development eliminated the need to work with radioactivity. The development of pulsed field gel-electrophoresis (Schwartz and Cantor, 1984), PCR (Mullis *et al.*, 1986) and the first automated DNA sequencing machine (Smith *et al.*, 1986) paved the way for DNA sequencing at the genomic level.

### 1.1.5 Genome Sequencing

The first genome to be completely sequenced was the genome of bacteriophage MS2 RNA, only 3,569 bases in length and containing just four genes (Fiers *et al.*, 1976). Just six years later, the comparatively much larger DNA genome of bacteriophage lambda, containing 48,502 bases and almost 100 genes was sequenced (Sanger *et al.*, 1982). The first free living organism to be completely sequenced was the bacteria *Haemophilus influenzae*, containing 1,830,137 bases and over 1,700 genes (Fleischmann *et al.*, 1995). The invention of shotgun sequencing (Venter *et al.*, 1996) accelerated the sequencing of the larger genomes of cellular organisms, and as of June 2005, 21 archaea, 207 bacteria, 33 eukaryota and over 1500 virus genomes have been sequenced, including higher eukaryotes such as human, mouse and rat (GOLD database, Bernal *et al.*, 2001). There are also numerous sequencing projects that sequence eukaryotic expressed RNA, these projects are fundamental to the development of more accurate gene prediction programs that can be trained with these expressed sequences in order to predict eukaryotic genes in genomic sequencing projects more accurately.

### 1.1.6 Sequence Databases

The International Nucleotide Sequence Database Collaboration consists of three sequence databases – GenBank (Benson *et al.*, 2005), EMBL (Kanz *et al.*, 2005) and DDBJ (www.ddbj.nig.ac.jp). All three databases exchange their sequence data on a daily basis, and as such contain virtually the same sequence data, but have different data formats. As of June 2005, GenBank release 148.0 contained 49,398,852,122 nucleotide bases from 45,236,251 sequences. Protein translations of these nucleotide sequences are also deposited in these databases. GenPept release 148.0 (translations from nucleotide files in GenBank) as of June 2005 contains 748,555,190 amino acids from 2,440,496 protein sequences. However, these sequence collections often contain

redundant entries, where entries may be truncated, identical, or contain small sequencing errors. In order to increase the quality and remove redundancy from protein sequence collections, several human curated databases have been developed. For example, release 11 (May 2005) of the RefSeq database (Pruitt *et al.*, 2005) at the NCBI contains 1,425,971 protein sequences. SWISSPROT (Bairoch and Apweiler, 2000; Boeckmann *et al.*, 2003) and PIR (Wu *et al.*, 2004) are also human curated databases. Redundant translations are removed and protein annotations are carefully verified. However, this process is time consuming, release 47.3 (June 2005) contains only 185,639 protein sequences. The TrEMBL (Boeckmann *et al.*, 2003) database complements SWISSPROT, and contains automatically generated translations of EMBL nucleotide sequences not yet included in SWISSPROT. Release 30.3 (June 2005) contains 1,782,502 protein sequences. The PIR database is descended from the first protein sequence database, the Atlas of Protein Sequence and Structure, created in 1965 (Dayhoff, 1965) and has been largely superseded by iProClass (Wu *et al.*, 2004) which consists of PIR, SWISSPROT and TrEMBL. Release 2.71 (June 2005) contains 1,891,813 protein sequences hierarchically classified into 36,000 PIR superfamilies and 145,300 families. In 2002, SWISSPROT, TrEMBL and PIR were merged into a single resource, the Universal Protein Resource (UniProt, Bairoch *et al.*, 2005).

### 1.1.7 Protein Structure Determination

Compared to the number of nucleotide and protein sequences, relatively few protein structures have been solved. All published protein structures are deposited into the Protein Data Bank (PDB, Bernstein *et al.*, 1977; Deshpande *et al.*, 2005). As of June 2005, the PDB contains 30,041 protein structures. 3,945 of these structures were determined by Nuclear Magnetic Resonance (NMR), the vast majority of the remainder were determined by x-ray crystallography.

X-ray crystallography determines the structure of a molecule from its diffraction pattern. X-rays are passed through a crystal containing a regular array of the molecule of interest producing a characteristic x-ray diffraction pattern, indicating the arrangement of molecules in the crystal. NMR can be performed on proteins in solution, and so avoids the crystallisation problems of x-ray crystallography. NMR exploits the phenomenon whereby some atoms will resonate when placed in a magnetic

field. In an NMR spectrometer this resonance is detected and amplified, the exact frequency of the resonance can be used to identify the type of atom. NMR analysis identifies distance constraints on residues that allow reconstruction of the underlying structure. NMR is restricted to smaller proteins, but can study protein conformational changes during protein folding or substrate binding.

Solving a protein structure is a time consuming and costly enterprise. The average cost for solving a single protein structure is about $250,000-$300,000. In an effort to affect a rapid increase in the number of solved protein structures, the National Institute of General Medical Sciences initiated a program in 2000 to solve 10,000 protein structures over ten years. This Protein Structure Initiative program, a collaboration between several structural genomic centres, has been running for five years. Around 1000 protein structures have been solved. The cost per protein structure was initially as much as $670,000, due mostly to high initial start-up costs. As these high-throughput centres progressed, the cost has dropped to around $180,000 and is expected to drop further to around $100,000 per protein structure as high-throughput pipelines are streamlined and become more efficient (Service, 2005).

### 1.1.8 Protein Structure

Proteins have a huge variety of functions and forms. Proteins can be characterised and compared in terms of protein structure. A protein's primary structure is the translated sequence of amino acids that form the peptide chain. The secondary structure of a protein is the local conformation of the peptide chain. The main units of secondary structure are the alpha helix and the beta strand. A typical helix ranges from as few as 5 to as many as 30 amino acid residues. The close packing of the helix minimises contact between hydrophobic carbon atoms and the surrounding water, and facilitates hydrogen bonding between amino acid amine group hydrogen atoms and carbonyl group oxygen atoms that stabilise the helical structure. 3.6 amino acid residues form a full turn of the alpha helix. A beta strand is also stabilised by hydrogen bonds, often several beta strands form a beta sheet. The beta strands hydrogen bond to each other to stabilise the beta sheet. As in the alpha helix, hydrophobic amino acid groups are packed on the internal face of a beta sheet, to minimise contact with water in the surrounding medium. Secondary structure elements interact with each other using

stabilising hydrophobic interactions. The tertiary structure of a protein describes the orientations of the secondary structures and their connectivity in three-dimensional space. Secondary structure elements can be closely packed in the tertiary structure. Helices can pack closely together to shield hydrophobic amino acid side chains, beta strands can pack together to form beta sheets, which in turn can pack against helices (shown in figure 1.3). In general the tertiary structure of a protein tends to pack hydrophobic amino acid residues internally, leaving hydrophilic amino acid residues accessible. Some proteins consist of more than a single peptide chain. Quaternary protein structure describes the relationship between different peptide chains of the same protein.

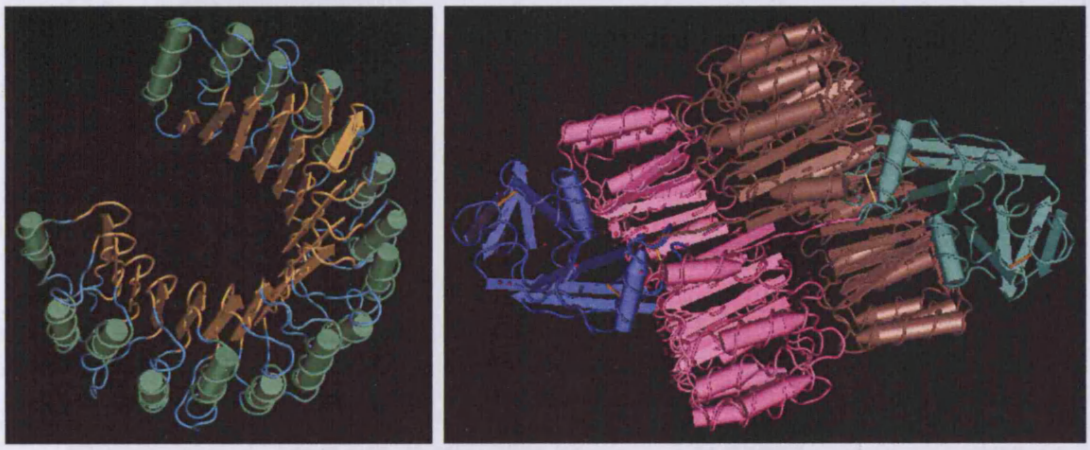

**Figure 1.3 Protein Structure.** *Tertiary structure (left) showing arrangement of alpha helices (green) and beta sheets (yellow) in a Ribonuclease Inhibitor domain. Quaternary structure (right) showing all four domains (blue, pink, brown, green) of Ribonuclease Inhibitor-Angiogenin Complex (PDB: 1A4Y, Papageorgiou et al., 1997).*

Many proteins form compact globular structures, some proteins appear to consist of distinct compact globular units linked together. These distinct protein subunits are termed domains. The exact definition of a domain ranges from 'a distinct globular protein subunit' to 'an evolutionary independent protein subunit'. These definitions can reflect both the physical properties of protein domains, as well as the evolutionary mechanisms by which proteins are thought to evolve. A protein domain may consist of a continuous region of a peptide chain, or may be formed by a discontinuous region of a peptide chain. A protein domain may even consist of more than one peptide chain.

### 1.1.9 Structure is more Conserved than Sequence

Chothia and Lesk (1986) first demonstrated the degree to which protein structure appears more conserved than sequence during evolution. This observation is reaffirmed by recent analyses of larger structural classifications (Orengo and Thornton, 2005), see figure 1.4 below.

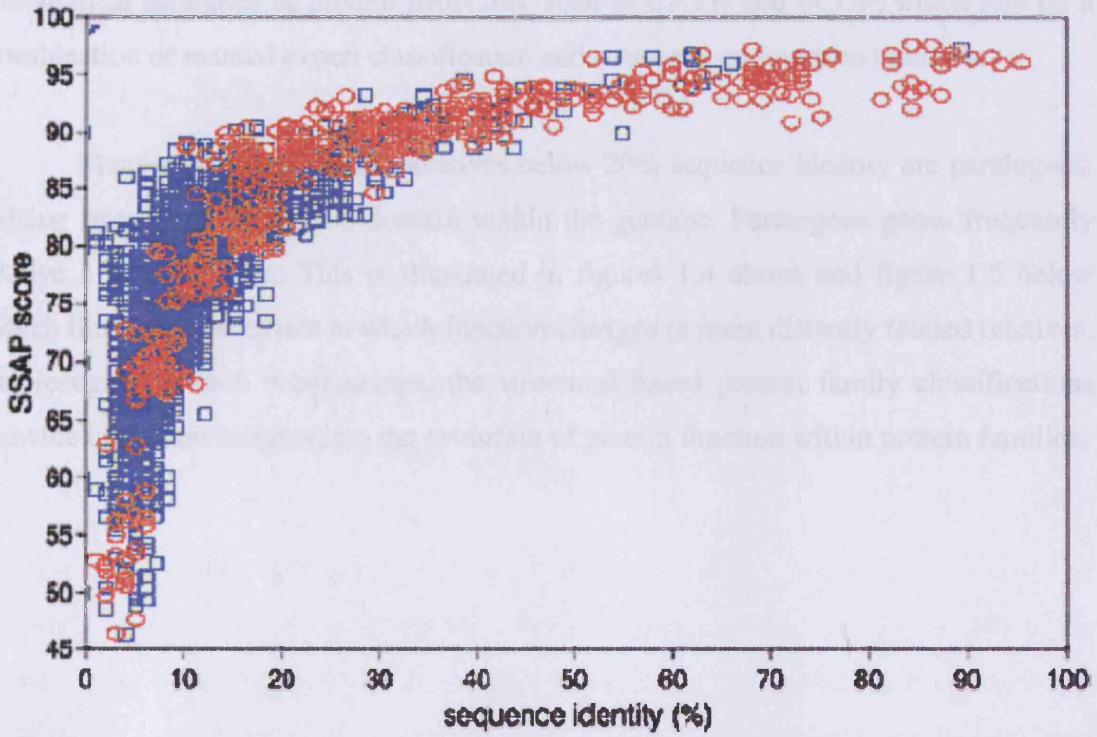


**Figure 1.4 Sequence Identity versus Structure Similarity.** *Sequence identity plotted against SSAP structural comparison score (from 0-100) for all pairs of homologous domain structures in the CATH domain structure database (red circles represent proteins with identical functions, blue squares represent proteins with different functions).*

Since structures are more highly conserved than sequences, structural similarity is more able to detect distant protein relatives than sequence similarity. Even with advances in sequence comparison methods, some remote homologues in the 'midnight zone' with less than 15% sequence identity can only be detected through protein structure comparison (Todd *et al.*, 2001; Orengo *et al.*, 2001). Structure based classifications that are able to incorporate these distant homologues provide protein

family datasets that permit further-reaching analyses of protein family evolution than sequence based resources alone.

The Protein Data Bank (PDB), based in the Research Collaboratory of Structural Biology (RCSB) Rutgers University, contains structures of over 30,000 proteins. These proteins are decomposed into over 60,000 protein domains of known structure. These structures can be clustered into protein families and superfamilies, producing hierarchical databases of protein structures, such as CATH and SCOP, which rely on a combination of manual expert classification and structural comparison methods.

Many of the very distant relatives below 20% sequence identity are paralogues, arising from duplication of a domain within the genome. Paralogous genes frequently evolve a new function. This is illustrated in figures 1.4 above and figure 1.5 below which illustrates the extent to which function changes in more distantly related relatives. By recognising such relationships, the structural based protein family classifications provide important insights into the evolution of protein function within protein families.


**Figure 1.5 Structure is more Conserved than Sequence.** *Functional diversity in distant homologous relatives in CATH. Whilst orthologues and paralogues are in the same Homologous Superfamily in the CATH database, the analogues have the same fold but are in different Homologous Superfamilies. Orthologues, paralogues and analogues can have very different sequence identities (green arrows) and functions (1tad – G1 subunit, Bos taurus; 1cip – G1 subunit, Rattus norvegicus; 1he8 – PI-3 kinase, Homo sapiens; 1n6h – Rab-5a kinase, Homo sapiens; 1e98 – Thymidylate kinase, Homo sapiens; 1srr – Sporulation response protein, Bacillus subtilis). Adapted from Orengo & Thornton, 2005.*

Function may be inherited at different levels of sequence identity with different degrees of confidence. For example Todd *et al.* (2001) analysed the relationship between EC number conservation and sequence identity and concluded that for single domain proteins, enzyme function as defined by the first three EC numbers is almost completely conserved between protein relatives which have a sequence identity of 40% or more, whereas in multi-domain proteins, the same level of conservation is seen between protein relatives which have a sequence identity of 60% or more (shown in

figure 1.6 below). Todd *et al.* conclude that this conservation of function allows functional prediction between protein relatives with at least 30% sequence identity with 95% accuracy in single domain proteins; whereas in multi-domain proteins below 40% sequence identity there is a significant reduction in conservation of function. This is in agreement with the analysis of Devos and Valencia (2000) who conclude that below 50% sequence identity, completely correct functional annotations cannot be inferred. However, Rost (2002) concludes that even at 50% sequence identity between proteins; complete EC number conservation only occurs in 30% of pairwise relationships.

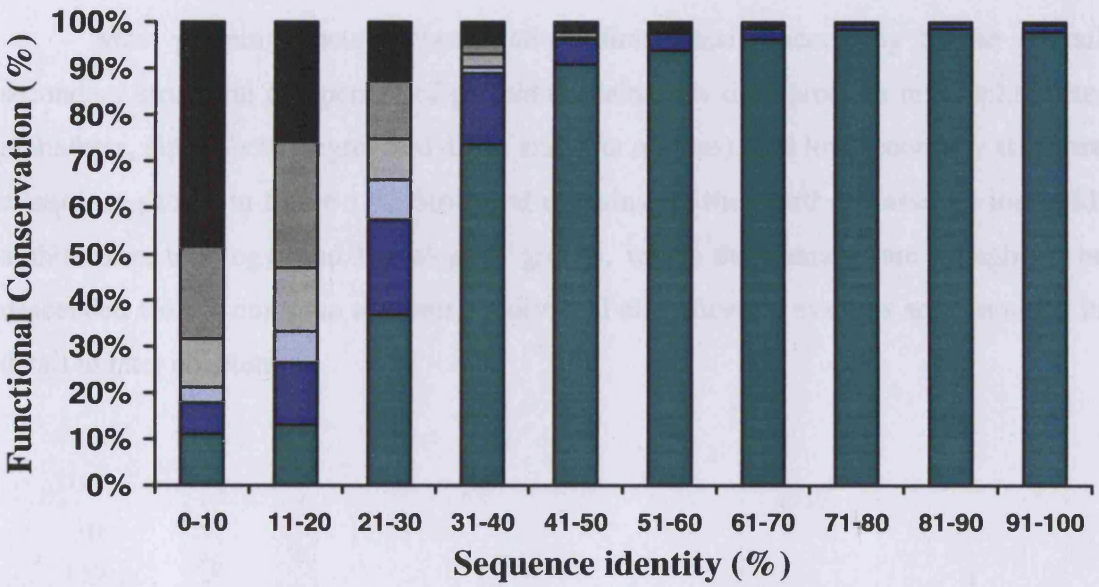


**Figure 1.6 Functional Conservation versus Sequence Identity in CATH.** *Conservation of functional annotation with sequence identity between pairs of related enzyme structures in CATH homologous superfamilies with at least two members. Percentage of total pairs conserved to four (green), three (blue), two (mauve), one (light grey) or no (dark grey) levels in the EC classification scheme. Pairs containing an un-annotated member are shown in black. Taken from Todd et al., 2001.*

### 1.1.10 Protein Structure Classification

Protein structures in the PDB are classified by several different resources (e.g. SCOP (Murzin *et al.*, 1995; Andreeva *et al.*, 2004); CATH (Orengo *et al.*, 1997; Pearl *et al.*, 2005); and the Dali domain dictionary (Holm and Sander, 1996; Dietmann *et al.*, 2001)). Most protein structure classifications initially deconstruct whole proteins into

protein domains. Some classification systems, such as the SCOP database, define protein domains as evolutionary conserved protein units. It has been estimated that around 65% of prokaryotic and up to 80% of eukaryotic proteins contain multiple domains (Apic *et al.*, 2001). Several algorithms have been written for recognising domains in protein structures (Jones *et al.*, 1998), often exploiting the fact that there are more contacts between amino acids within a domain than between different domains, or searching for hydrophobic clusters that could represent domain cores (DETECTIVE, Swindells, 1995). However, most classifications rely on manual intervention to some extent to determine domain boundaries.

Most protein structure classifications first classify according to the overall secondary structural component of protein domains, dividing proteins into alpha, beta, alpha/beta, alpha+beta (segregated alpha and beta regions), and low secondary structure classes, as shown in figure 1.7. Structural domains are then further classified into fold, architecture, topology, and homologous groups, where the domains are thought to be descended from a common ancestor. Individual classification systems are discussed in detail in later chapters.

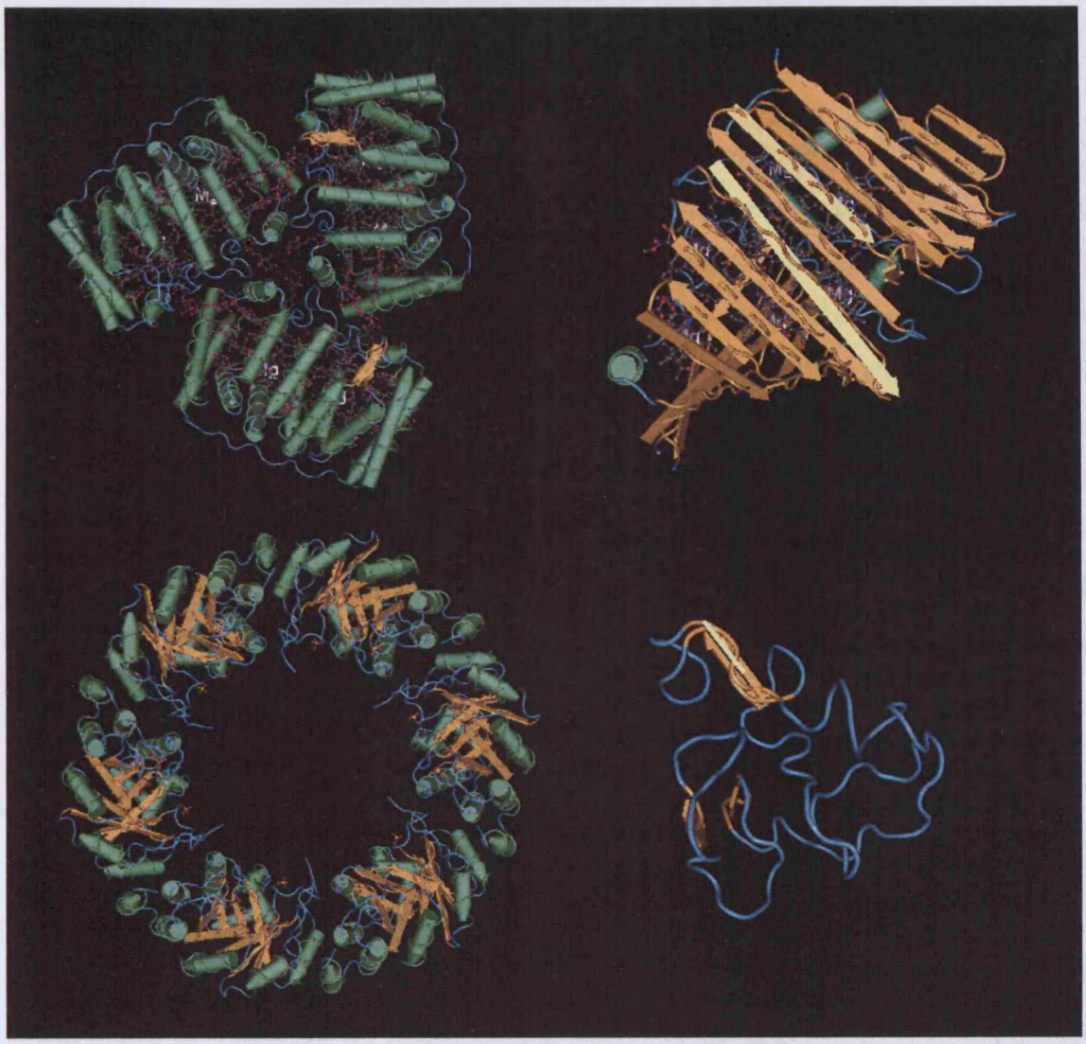

**Figure 1.7 Protein Structure Classification.** *Secondary structure denomination of proteins into Alpha (top left, PDB:1ppr, Hofmann et al., 1996), Beta (top right PDB:4bcl, Tronrud and Matthews, 1993), Alpha/Beta (bottom left PDB:1g61, Groft et al., 2000) and Low secondary structures (bottom right PDB: 1jfw, Peloponese et al., 2000) classes.*

## 1.2 Homology

Homology means a common origin can be inferred between two entities. In protein terms, homologous proteins are proteins thought to have a common evolutionary ancestor. Organising sequences and structures into homologous groups not only helps to classify the ever increasing amounts of biological data presently being generated, but has practical implications for the study of biological relationships within and between families of homologous proteins. Statistically high levels of similarity between proteins can be used to infer homology between them. Methods of detecting homology between proteins inferred from protein sequence similarity and protein structure similarity are described below. Context based methods that use genomic information to infer functional relationships between proteins that may indicate a linked evolutionary history, but not necessarily a common evolutionary ancestry, are also described.

### 1.2.1 Sequence Based Homology Detection Methods

Pairwise sequence similarity methods can be divided into local similarity methods and global similarity methods. Local similarity algorithms (for example Smith and Waterman, 1981) identify conserved regions between two sequences and ignore regions with little or no similarity, whereas global similarity algorithms (for example Needleman and Wunsch, 1970) optimise the overall alignment of two sequences and include regions with little or no similarity. Sequences are aligned and scored according to sequence identities (identical residue pairs), acceptable substitutions, and gaps in the alignment. However, there may be very many different possible alignments between long sequences and scoring each alignment could be computationally intensive. This problem is solved using dynamic programming. Dynamic programming is a technique that divides the alignment problem into stages. The initial stage creates a matrix grid where the residues in one sequence run along the x-axis, and the residues in the other sequence run along the y-axis. A scoring matrix is generated, whereby each square in the matrix is sequentially filled, by calculating a running score of the similarity between the residues leading to that square. The last stage traces a path back through the scoring matrix to find the optimum scoring path. This path through the matrix can be converted into an alignment between the two sequences. For global alignments, the entire path

through the matrix is calculated, for local alignments, only a subsequence region corresponding to the highest partial score need be reported.

#### 1.2.1.1 BLAST1

For searching large sequence databases for sequence relatives, aligning all the sequences in the database to the query sequence using dynamic programming is still too time consuming. Heuristic approaches that approximate the scoring matrix and optimal path can speed this process. The BLAST1 algorithm (Altschul *et al.*, 1990) shortens the search time for an optimal path by looking for non-gapped alignments. The query sequence is broken into word fragments (default word size is four), each word is scored against a substitution matrix (PAM120, Dayhoff *et al.*, 1978) and all substitutions for each word that score above a threshold are then used to scan against each sequence in the sequence database. Every time a word matches a database sequence, the word alignment, or maximal segment pair (MSP), is extended at either end and scored. The highest scoring MSP (known as the HSP) is then returned. The FASTA algorithm (Lipman and Pearson, 1985) is similar to BLAST1 in that it also sacrifices precision for speed. FASTA looks for exact matches between short sequence regions. If enough short regions of identity are found, FASTA uses dynamic programming to find the optimal path through the matrix. The size of the exact matches determines the speed and accuracy of the method. The longer the size requirement, the faster an optimal path can be calculated, but the less likely the optimal path is to contain such a long exactly matching region.

#### 1.2.1.2 BLAST2

Both FASTA and BLAST1 look for non-gapped alignments. BLAST2 (Altschul *et al.*, 1997) is a modification to BLAST1 that permits gapped alignments (Gapped-BLAST) and iterative searches (PSI-BLAST). BLAST2 uses a two-hit method, whereby two words need to be found within a threshold distance of each other before word extension is permitted. This dramatically reduces computational time as fewer word extensions need to be performed. Permitting gapped alignments reduces the number of word searches of database sequences, since a word alignment is more likely

to be extended further since gaps in the alignment are permitted. Successive iterations of sequence searching and alignment produce position specific scoring matrices (PSSM's) that reflect conserved sequence regions, whilst allowing sequence variation in non-conserved sequence regions. This can be used to score successive iterations of sequence searches. This iterative searching method makes BLAST2 more sensitive in detecting sequence relatives with low sequence similarity than a single iteration search. The position specific scoring matrix (PSSM) generated by BLAST2 describes the proteins that align with significant scores to the query sequence during the iterative process. In sequence databases containing families of related sequences, a PSSM can be generated to profile each family. A library of PSSMs can then be quickly searched to identify query matches to each family without having to search the entire sequence database (IMPALA, Schaffer *et al.*, 1999).

### 1.2.1.3 Expectation Values

The significance of any pairwise alignment between two sequences can be quantified using an expectation value (E-value), which gives an indication of the likelihood of an alignment score (S) occurring by chance. An E-value is determined by multiplying the size of the database being searched by the p-value. In an alignment of two proteins of length m and n (where length is determined by the number of amino acid residues in each protein sequence), the p-value is calculated from the equation: p-value = $Kmne^{-lambdaS}$. Where the parameter $Ke^{-lambdaS}$ represents the probability of an HSP with score S occurring by chance. This value is pre-calculated by fitting the tail of the distribution of scores returned by random, unrelated sequences (which produces an extreme value distribution in which the tail decays more slowly than for a normal distribution). K and lambda are empirically derived parameters, the values of which are functions of the alignment scoring matrix used, and can be thought of as natural scales for the search space size and the scoring system respectively. When searching databases using BLAST, the size of the database being searched is represented in the parameter n, calculated as the total length of all sequences in the database, effectively considering the database sequences being searched against as one long single sequence. Thus the the database size parameter n is multiplied by the pairwise p-value to produce a BLAST E-value in the equation above. This generates an E-value that decreases exponentially with score. BLAST E-values do not refer to the whole query sequences,

but rather to the fragment of the query sequence that is given in the match (i.e. the HSP). A useful guide to E-values can be found at the NCBI (http://www.ncbi.nlm.nih.gov).

#### 1.2.1.4 Hidden Markov Models

Profiles generated from multiple sequences better describe conserved and variant sequence regions. PSI-BLAST and IMPALA profile libraries are capable of detecting distant, yet significant sequence similarity proteins that pairwise methods like BLAST1 are unable to distinguish. More recent developments of profile methods include Hidden Markov models (HMMs, Eddy 1998). A Markov model describes a set of 'states', each state being defined by a probability distribution. Markov models are subject to three assumptions: the next state is dependent only on the previous state(s); state transition probabilities are independent of time; and observed state outputs are statistically independent of previous outputs. State transition probabilities can be described in a matrix describing all possible transition probabilities between each state in the model.

An HMM describes two sets of states that are closely linked, one set of known states and another set of unknown, or hidden states. In protein sequence terms, each observed state represents an amino acid in a protein family. Known states are the observed amino acid frequencies at a given position in the model, and unknown states are the mechanisms and processes responsible for the observed sequence variation. Unlike other profile methods, HMMs can be built from unaligned protein families. In addition to a state transition matrix, an HMM is enhanced by a confusion matrix, describing the probability of observing a particular known state given that the hidden model is in a particular hidden state.

A linear HMM for a protein family is a model describing observed states corresponding to columns in a multiple alignment (shown in figure 1.8). The transition and confusion matrices are generated from the observed amino acid residues of protein family members. Each observed state has three hidden states that model matches, insertions, and deletions. This allows an HMM to model position dependent amino acid distributions and position dependent insertion and deletions (Krogh *et al.*, 1994).

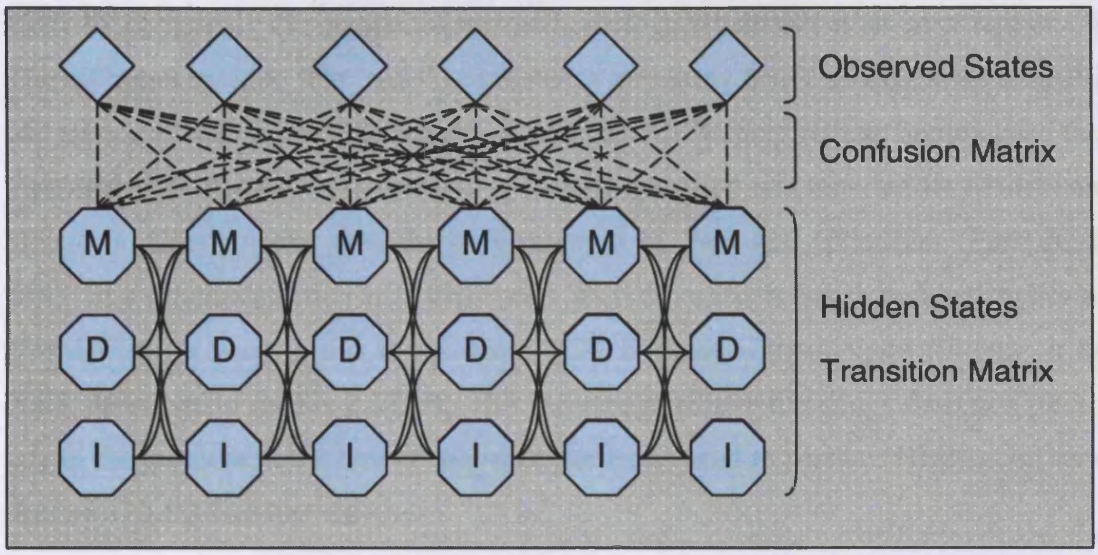


**Figure 1.8 Linear HMM.** *Amino acid residues in family members are represented by observed states (diamonds). The transition matrix (full lines) describe the probability of traversing between each hidden state (octagons) based on the previous state. The relationship between observed and hidden states M (match), D (deletion) and I (insertion) is described by the confusion matrix (dashed lines).*

Query sequences are searched against libraries of HMMs representing protein families, or protein domains. Query sequences are scored against each HMM by calculating the probability of the observed amino acid at each model state. The optimum path through an HMM can be identified in a similar way to finding the optimum path through a scoring matrix, by using the Viterbi algorithm (Viterbi, 1967). The Viterbi algorithm can be used to calculate both local and global optimum paths.

### 1.2.1.5 Comparison of BLAST1, BLAST2 and HMMs

There are two main software packages available for building and scanning HMMs. SAMT (Karplus *et al.*, 1998) and HMMER (Eddy, 1998) both use linear HMMs. These packages have been compared to each other and BLAST1/BLAST2 by Madera and Gough (Madera and Gough, 2002) using an all against all search of 2873 domains of known structure and less than 40% sequence identity, where the evolutionary relationships between the domains are known (a total of 36 612 possible

true homologous pairwise relationships). Comparisons between each method were made by comparing the number of remote homologues detected at an error rate of 1%. The authors conclude that SAMT consistently produces better models than HMMER and detected 24% of possible remote homologues, 10% more remote homologues than HMMER and BLAST2. The remote homology detection rate reported by Madera and Gough is slightly lower than the 34% reported by Park and colleagues (Park *et al.*, 1998) in a similar detection test. Four years later, more recent benchmarking studies for HMM libraries used in this thesis detect 76% of remote homologues (Sillitoe *et al.*, 2005; described in section 2.4.2.2). This increase in remote homology detection rates is due to the increase in the size of sequence databases used to produce HMMs and more sensitive HMM building methods.

### 1.2.2 Structure Based Homology Detection Methods

Proteins sharing extremely low levels of sequence similarity can still possess a high level of structural similarity, from which homology can be inferred, see example in figure 1.9. Studies of protein structural families have shown that homologues share on average only 15% sequence identity (Todd *et al.*, 2001).

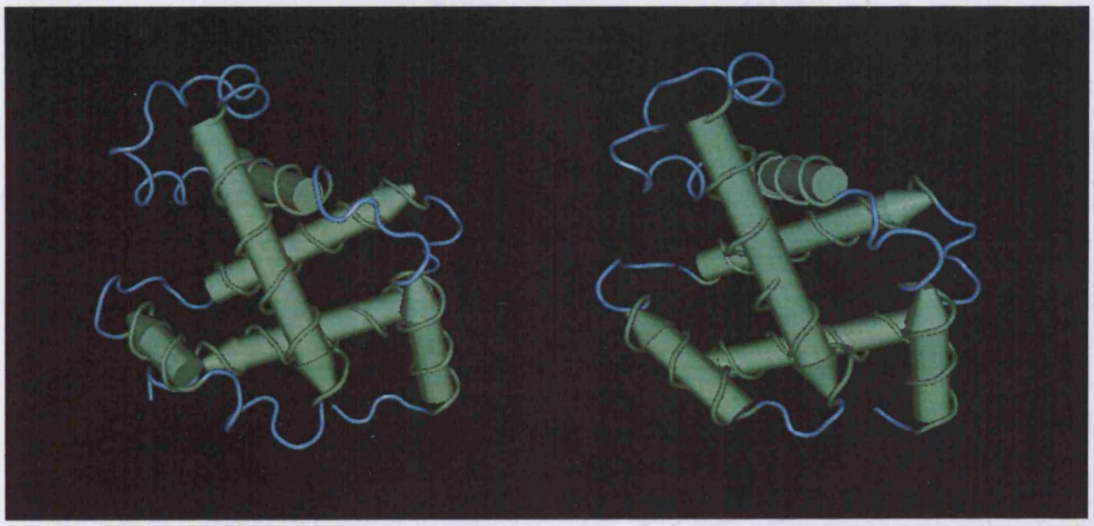

**Figure 1.9 Structural Similarity in the Absence of Sequence Similarity.** *The structure of human haemoglobin (left, PDB: 1hho, chain A, Shaanan, 1983) and insect haemoglobin (right, PDB: 1eco, Steigemann and Weber, 1979) show similar 97 amino acid core structure (cRMS=2.08A), with a sequence identity of 12.4%.*

### 1.2.2.1 Structural Alignment

Structure comparison and alignment algorithms were first introduced in the early 1970s. The rigid body superposition methods (Rossmann and Argos, 1976) allow structures to be superposed and a similarity measure calculated. Structures are compared by superposition of equivalent peptide chain carbon-alpha atoms, or whole secondary structures. Structures are translated and rotated relative to one other until the difference between putative equivalent residues or secondary structure elements is minimised. Similarity can be scored by calculating the root mean square distance (RMSD) between equivalent peptide chain carbon-alpha atoms (cRMS), residues or secondary structure elements. Superposition methods are limited when aligning distant homologues that may contain large insertions and deletions or changes in the orientations of equivalent secondary structures. More sensitive alignment algorithms based on dynamic programming, secondary structure alignment and fragment comparison have been developed. Many protein structure classifications use secondary structure based comparison (GRATH, Harrison *et al.* 2003; SEA, Rufino and Blundell 1994) to identify putative structural relatives and then apply slower, more accurate residue based methods. More computationally intensive residue based comparisons (COMPARER, Sali and Blundell 1990; SSAP, Taylor and Orengo 1989; STAMP, Russell and Barton, 1992; DALI, Holm and Sander, 1993; CE, Shindyalov and Bourne 1998) result in accurate structural alignments. Rather than attempting to superpose equivalent residues between protein structures, many of these methods compare the internal distances between residues within the same structure to align residues with similar sets of internal distances.

### 1.2.2.2 Threading

Threading techniques (for example, Jones *et al.*, 1992; Bryant and Lawrence, 1993) attempt to position a protein sequence onto a structural template by identifying the most energetically stable fit to the template. Threading techniques require a database of known structural templates to scan against, an energy function for measurement of the energy of the protein sequence–structural template alignment, an

algorithm for identifying the optimal alignment and finally a scoring system to measure the reliability of the structural prediction. Threading assumes that since many different protein sequences are known to fold into a limited number of protein folds, by attempting to thread a protein sequence into each member of a known structural template library, protein sequences can be assigned a structural fold. Threading scores the likelihood of two residues occurring at a certain distance relative to each other in space for each residue pair in the query-template alignment. Because threading techniques also incorporate some structural data, they are used for predictions at very low sequence identities (<25%) where sequence-based homology methods may not function. The performance of different threading approaches is assessed every two years in CASP fold recognition predictions (Kinch *et al.*, 2003). Threading is computationally expensive and most threading-based methods are too slow for large-scale genomic annotation projects (Cherkasov and Jones, 2004). However, some annotation protocols incorporate threading potentials. For example, GenThreader (McGuffin and Jones, 2003) consists of a neural network trained to combine BLAST2 alignment profiles (seeded with structural alignments), secondary structure predictions and energy potentials derived from threading.

### 1.2.3 Context Based Functional Prediction Methods

Implication of a functional linkage between proteins can be evidenced by methods other than sequence or structure similarity. A variety of different methods have been developed to describe contextual relationships between proteins. Most of these methods rely on genomic information to infer a functional linkage between proteins.

#### 1.2.3.1 Rosetta Stone

The Rosetta Stone method identifies homology between proteins when homologues are found fused together in another organism. If two proteins are located apart in one genome, but can be identified fused together in another genome, homology between them can be implied. Eisenberg *et al.* (Eisenberg *et al.*, 2000) cite the example of yeast proteins involved in tryptophan biosynthesis, TrpG and TrpF, the *Escherichia*

*coli* homologues of which are found fused together in a tryptophan biosynthesis single protein, TrpC. Other examples include *Caenorhabditis elegans* protein Ade5,7,8 involved in the biosynthesis of purines, which is homologous to two separate protein sequences, Pur2 and Pur3 in the yeast genome that perform the same function.

### 1.2.3.2 Protein-Protein Interaction

Proteins that interact physically are functionally linked in that both proteins are likely to be involved in similar functions. Homologous sequences are likely to share similar interactions. Evidence of physical interactions between proteins can be obtained from high-throughput technologies such as large-scale two-hybrid screens, protein microarrays, and mass spectrometry of protein complexes. However, the interactions identified from these multiple experiments can be contradictory (Mrowka *et al.*, 2001), requiring development of methods to assess the reliability of protein-protein interaction data (von Mering *et al.*, 2002, Saito *et al.*, 2003).

### 1.2.3.3 Synteny

Especially in prokaryote genomes, genes that are in close proximity to one another in several genomes are likely to be functionally linked (Dandekar *et al.*, 1998). Such linkage has been demonstrated by genes in the purine biosynthetic pathway, tryptophan biosynthesis, glycolysis and signal transduction pathways (Overbeek *et al.*, 1999). Studies indicate that a minimum of 10 genomes are required in order to detect even small functional clusters by conservation of gene proximity alone (Overbeek *et al.*, 1999), since gene proximity is not often conserved above proximal gene pairs of proteins (Bansal, 1999). Gene proximity correlations have been found to a lesser extent in eukaryotes (Barbazuk *et al.*, 2000). Gene proximity can be combined with other techniques to detect more diffuse but significant clusters of genes showing functional linkage (Kolesov *et al.*, 2001).

### 1.2.3.4 Phylogenetic Profiles

The presence or absence of a protein homologue in multiple genomes is called a phylogenetic profile. Eisenberg *et al.* (Eisenberg *et al.*, 2000) note that when considering presence or absence profiles, there are far more possible phylogenetic profiles than protein families, so that the phylogenetic profile of a protein is an almost unique pattern of its genomic distribution, and any proteins with similar profiles are likely to be functionally linked. Phylogenetic profile comparison has been used to identify proteins functionally related to ribosomal proteins (Pellegrini *et al.*, 1999), relationships that are not detectable by sequence comparison. Phylogenetic profiles consisting of a range of values, rather than the simpler presence/absence profiles are able to discern more subtle profile relationships (Date and Marcotte, 2003).

### 1.2.3.5 Expression Profiles

High-throughput microarray protein expression experiments allow proteins to be associated by co-expression profile analysis. Proteins that are found to have correlated expression profiles are likely to be functionally related. For example, Baldessari *et al.* (Baldessari *et al.*, 2005), describe the identification of thirteen groups of functionally related synergistically expressed proteins involved in diverse molecular processes including RNA processing, cell cycle, respiratory chain and protein biosynthesis, in a large scale microarray analysis of gene expression in *Xenopus laevis*. Analysis and comparison of expression profiles across multiple microarray experiments is often complicated by the range of different statistical normalisation and clustering techniques, experimental conditions and different microarray platforms (Cope *et al.*, 2004; McShane *et al.*, 2002; Pavlidis and Noble, 2001).

## 1.3 Functional Annotation

The different functional roles performed by different proteins can be characterised experimentally. However, only a small number of proteins have had their function determined experimentally, the vast majority of proteins are functionally annotated by inheriting annotations from functionally characterised homologues. Functional annotation has been shown to be reliably inherited between proteins with as little as 30-40% sequence identity (Hegyi and Gerstein, 2001). However, Todd *et al.* (Todd *et al.*, 2001) document examples of proteins with very low sequence identity sharing similar functions (for example chymotrypsin and subtilisin) and similar proteins with different functions (for example lactalbumen and lysozyme). However, exhaustive analysis over a large dataset showed that typically one needs 40% sequence identity in single domain proteins (60% sequence identity in multidomain proteins) to reliably inherit function at 95% confidence.

### 1.3.1 Defining Function

Different experimental approaches to determine protein function identify different kinds of functional associations between proteins. Protein expression experiments determine cellular process functions whereas protein-protein interaction experiments define a molecular association between proteins. Whilst cell biologists might use cellular processes to define protein function, molecular biologists may use molecular chemistry to define a protein's function. Functional databases try to formalise and standardise functional terminology. Many functional ontologies employ a hierarchical classification of function, for example the Enzyme Commission (Webb, 1992). Some protein functions are difficult to classify in a linear hierarchy: multifunctional proteins require linkages between hierarchical elements of the ontology, and different functional descriptors cannot be integrated into a single hierarchy. The Genome Ontology (GO, Ashburner *et al.*, 2000) tries to overcome these problems in two ways. Firstly, protein functions are defined by three separate ontologies. Each defines a different kind of functional descriptor: molecular function, biological process and cellular component. These three ontologies allow different kinds of functional terms to be assigned to the same protein. And secondly, instead of a linear hierarchical classification scheme, GO functions are organised into directed acyclic graphs, allowing

multiple linkages between hierarchical terms. This data structure allows classification and comparison of proteins with multiple functional terms or capable of multiple functions. Some functional resources are described below.

### 1.3.2 Enzyme Commission

The Enzyme Commission (Webb, 1992), was initiated in 1955 by the International Union of Biochemistry (IUB) in consultation with the International Union of Pure and Applied Chemistry (IUPAC) with the intent 'to consider the classification and nomenclature of enzymes and coenzymes, their units of activity and standard methods of assay, together with the symbols used in the description of enzyme kinetics'. The resulting classification scheme, EC Numbers, is a hierarchical classification of enzyme catalysed reactions, consisting of four numbers for each hierarchical level. The highest level consists of six groups: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. The next two numbers indicate the actor and acceptor molecular groups involved in the reaction. The last number specifies the substrate.

### 1.3.3 Kyoto Encyclopaedia of Genes and Genomes

The Kyoto Encyclopaedia of Genes and Genomes (KEGG, Kanehisa *et al.*, 2004) is a collection of databases including descriptions of biological processes (PATHWAY database), chemical reactions (LIGAND databases) and gene/protein sequence data (GENES database).

### 1.3.4 Clusters of Orthologous Groups

The Clusters of Orthologous Groups (COG, Tatusov *et al.*, 2003) database contains 4873 groups of protein orthologues from 66 completely sequenced prokaryotic and unicellular organisms. The database also contains 4852 clusters of orthologues from 7 eukaryotic genomes (KOG). Clusters are defined by identification of genome-specific best hits between proteins using all against all BLAST2 comparison. COGs and KOGs are divided into 23 broad functional process classes (the largest of which

include: E - amino acid metabolism and transport, C – energy production and conversion, J – translation, G – carbohydrate metabolism and transport, L – replication and repair, and T – signal transduction) obtained from GenBank and other public databases and by searching primary literature. Functional annotation of COGs and KOGs is undertaken on a case by case basis, using published data, protein domain analysis, phyletic patterns and gene order conservation.

### 1.3.5 Genome Ontology

The Genome Ontology (Ashburner *et al.*, 2000) was initiated by eukaryotic model organism databases (FlyBase, Saccharomyces Genome Database and the Mouse Genome Database) in 1988. It now includes many databases representing prokaryotes and eukaryotes. GO consists of three separate hierarchical ontologies describing molecular functions, biological processes and cellular component. GO ontologies are organised into directed acyclic graphs where a child term may have multiple parent terms.

### 1.3.6 Affymetrix

Affymetrix is a manufacturer of microarrays. The company provides annotation for each gene represented on each microarray. This annotation is derived from public databases and includes gene identifiers and description from the NCBI, domain classifications from SCOP and Pfam, EC Numbers and KEGG annotations. Affymetrix microarrays are used by many different research groups. This permits expression data deposited in public databases (for example the NCBI Gene Expression Omnibus, Barrett *et al.*, 2005) to be collated between different microarray experiments from multiple sources.

### 1.3.7 STRING

STRING (von Mering *et al.*, 2005) contains predicted and known protein-protein interaction data. These interactions may be physically interacting proteins or

functionally linked proteins via linkages inferred through genomic context, microarray co-expression experiments, and the COG and KEGG databases. STRING currently contains 730,000 proteins from 180 genomes.

### 1.3.8 Reliability of Annotation

Experimentally determined data can produce remarkably little consistency between different experimental methods. For example, in two large-scale mass spectrometry protein-protein interaction studies in yeast, only 19.2% and 27.5% of the interactions identified using each method were identified by both methods (von Mering *et al.*, 2002). Aside from errors caused by the inconsistency of experimental methods, assigning annotation to proteins by inheriting annotations from related proteins can cause annotation quality problems (Bork and Koonin, 1998). In smaller databases curated by experts, annotation is likely to be manually checked and verified, where possible, against the literature. However, since few annotations are directly experimentally characterised, the majority of databases contain large amounts of inherited annotation (Karp *et al.*, 2001). The mechanism of annotation by inheritance, by its very nature leads to error propagation. The protection given to sequence annotations in primary databases (whereby only the submitter of the sequence can alter it's annotation) means that it is difficult to correct errors once identified, and that errors persist. Even once corrections to the original entry are completed, erroneously annotated homologues may still persist. There are several annotation analysis programs that aim to identify erroneous annotations in database collections (Wieser *et al.*, 2004; Kaplan and Linial, 2005), whilst other approaches focus on assigning reliability scores to annotations (Valencia, 2005).

## 1.4 Objectives

The following five chapters divide this research into different sections; but all these chapters are primarily concerned with describing the distribution and relationships between protein families and domain families in completely sequenced genomes. Chapter two describes the construction of Gene3D, a resource containing completely sequenced genomes that are annotated with protein family, domain family and functional information. Chapter three describes the distribution of protein families and domain families across these completed genomes. Chapter four is concerned with how well the domain families identified in the genomes have been characterised, and how many more families would be required to be better characterised in order to more accurately describe the domain complement of these completed genomes. Chapter five is divided into two main sections. The first section describes the distinct distributions of domain families identified within bacterial genomes. The second section introduces a novel phylogenetic profile method that can be used to infer evolutionary and functional relationships between domain families. Finally, chapter six summarises the conclusions of this thesis, and suggests future work to improve and expand these analyses.

# CHAPTER TWO

# Construction of the Gene3D Resource of Complete Genomes Annotated with Protein Family, Domain Family and Functional Information

## 2.1 Introduction

### 2.1.1 The Repertoire of Completed Genomes

The advent of completely sequenced genomes permits the complete protein component of certain organisms to be identified and opens new avenues of research in the biological sciences. Currently, the Genomes On-Line Database (GOLD, Bernal *et al.*, 2001) reports 261 published completely sequenced genomes (plus an additional ~1500 viral genomes) and over 1000 ongoing genome sequencing projects (as of June 2005). The number of completely sequenced genomes has risen steadily and is currently increasing at a rate of about one per week (shown in figure 2.0).

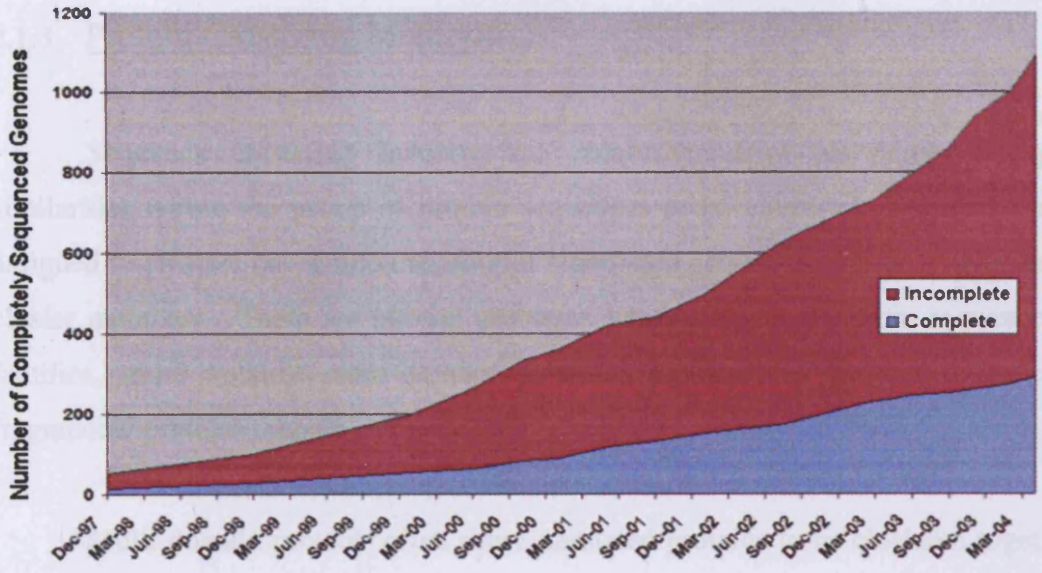


**Figure 2.0 Increase in Genomic Data.** *Increase in the number of completely sequenced genomes and the number of ongoing sequencing projects. Taken from the Genomes On-Line Database.*

Completely sequenced genomes from all three kingdoms of life are available. There are currently 33 complete eukaryotic genomes (including chromosomes) ranging from single cellular yeasts to multicellular fungi, plants and animals; 207 complete bacterial genomes; and 21 complete archaeal genomes. The proteins encoded by these genomes may be clustered into families and decomposed into domains. Analysis of the distribution of these families and domains and of the combinations of domains that are found can tell us much about the evolution of genomes at the molecular level (Vogel *et al.*, 2004).

### 2.1.2 Protein Annotation

The massive increase in protein sequence databases (for example, GenBank grows exponentially, currently doubling in size every 12 to 15 months, NCBI News, Summer/Fall 2004 edition) requires fast, automated methods of protein annotation. Protein clustering into families of related proteins permits inheritance of annotation from a consensus of multiple related proteins, which can be more reliable than from individual pairwise comparison (Devos and Valencia, 2000). These approaches and technologies are discussed below.

### 2.1.3 Protein Clustering Methods

Sequence clustering involves the measurement of all pairwise sequence similarities within the group of protein sequences to be clustered. Proteins are then assigned to clusters based upon sharing of significant sequence similarity with existing cluster members. There are several problems when clustering protein sequences into families, most notably multi-domain proteins, promiscuous protein domains and fragmented proteins (shown in figure 2.1).

Multi-domain proteins often cause unrelated proteins to be clustered together, if they share a significant proportion of similar domains (Doolitte, 1995; Smith and Zhang, 1997). Detection of shared protein domains does not necessarily indicate that the proteins have a biochemical function in common (Henikoff *et al.*, 1997), unless the domain context of the proteins is shared (Hegyi and Gerstein, 2001). Promiscuous

domains, for example SH2, WD40 and DnaJ domains that occur with high frequency in many proteins with different functions can produce significant sequence similarity between otherwise unrelated proteins. Fragmented proteins, incomplete protein sequences in protein databases, can render accurate protein domain determination by sequence comparison methods unreliable and incomplete. Some clustering methods attempt to overcome these problems by identifying protein domains within multi-domain proteins, for example by using BLAST reports (Guan, 1997), domain libraries (Pfam – Bateman *et al.*, 2004; ProDom – Servant *et al.*, 2002) and iterative sequence comparisons (GeneRAGE – Enright *et al.*, 2000). However some of these methods still require manual intervention to cluster multi-domain proteins correctly or are too computationally intensive for use on large protein datasets. Of the more recent clustering methods, TribeMCL (Enright *et al.*, 2002), ADDA (Heger and Holm, 2003) and CHOP (Liu and Rost, 2004), described below, are notably able to work efficiently and reasonably accurately with very large data sets (e.g. SWISSPROT/TrEMBL) and so make it feasible to cluster large collections of complete genomes.

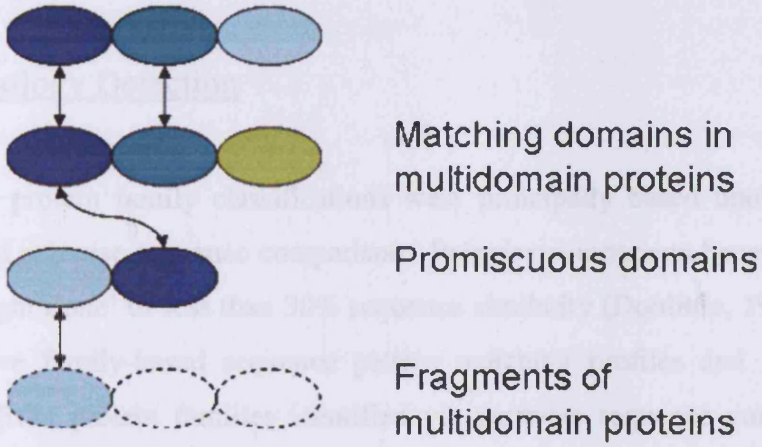


**Figure 2.1 Problems Associated with Clustering.** *Matching domains in multidomain proteins, promiscuous domains and fragments of multidomain proteins can erroneously indicate protein relatedness between otherwise unrelated proteins via regions with sequence similarity (arrows).*

2.1.3.1 Subclustering Families

Subclustering within families on the basis of sequence identity allows identification of more closely related relatives within a family. There are three main

approaches: single linkage, multi-linkage and directed multi-linkage clustering, which can produce very different subclusters within a protein family.

Using the same sequence dataset, three different approaches yield different clusters. Single linkage clustering permits membership of a cluster when significant sequence similarity exists to any existing cluster member, giving more diffuse clusters since not all cluster members are required to have significant sequence similarity to each other. Multi-linkage clustering however, requires that all existing cluster members share significant sequence identity to each other, producing smaller, tighter clusters. Multilinkage clustering can be dependent on the order on which sequences are clustered; a change in the order can produce different clusters. Directed multi-linkage clustering is based upon the same principle as multi-linkage clustering but in addition, clusters are ordered by descending similarity; the most similar sequences are clustered together first. Directed clustering is thus not effected by the order in which sequences are added to clusters.

### 2.1.4 Homology Detection

Early protein family classifications were principally based upon detection of homology via pairwise sequence comparison. Pairwise comparison becomes unreliable in the 'Twilight Zone' of less than 30% sequence similarity (Doolittle, 1990), however, more sensitive family-based sequence pattern matching profiles and motifs can be constructed from protein families identified via pairwise sequence comparisons (for example PRINTS, Attwood *et al.*, 2003). Sequence profiles and motifs often represent highly conserved residue signatures within a protein sequence that may be associated with a particular evolutionary family or biological function. Using these profile methods can often detect homologues at low levels of sequence identity where pairwise comparison becomes unreliable.

Profile methods representing whole protein sequences, for example PSI-BLAST, are able to describe position specific probabilities of protein residue insertions and deletions occurring within families. The most recent development in profile methods are hidden Markov models. These models are better at modelling insertions and

deletions and are capable of detecting more remote homologues in large databases than previous methods (described previously, see section 1.2.1).

#### 2.1.4.1 Building Hidden Markov Models using SAMT

There are two well established HMM program suites, HMMER and SAMT (described previously, see section 1.2.1.5). Both these program suites allow building of HMM libraries and searching databases with HMM libraries. At the time of starting this project, the SAMT program suite was used since it had been shown to be more powerful than HMMER in detection of remote homologues and at the time was slightly faster to search databases with HMM libraries (James Bray, PhD thesis). In the SAMT HMM software package (Karplus *et al.*, 1998) model building process (shown in figure 2.2), an initial BLAST is performed to identify a set of close relatives and a set of more distant relatives. The sequence set containing close sequence relatives is used to generate an initial sequence alignment for the seed sequence. An initial model is produced to represent the states in this alignment. This model is used to score all the sequences in the sequence set of distant relatives. In this first iteration of scoring, a very stringent E-value cut-off is used. Sequences in the distant relatives sequence set that score below this cut-off are pulled into the initial alignment to produce a first iteration alignment.

In the second iteration and in subsequent iterations of the model building process, the E-value cut-off for inclusion of new sequences into the alignment is successively increased. Increasing the E-value cut-off allows more distant sequence relatives to be gradually incorporated into the alignment and thus the model. New sequences are given less weighting in the alignment and model building process so that the sequence alignment signature of the close relatives from the initial iteration is preserved, and yet the sequence diversity of distant relatives is expanded in the model. After the sixth iteration the resulting model is the HMM representing the seed sequence family. The SAMT program suite allows the HMM model building process to be subtly altered by changing many different variables including the number of iterations, the cut-off E-values at each iteration and the weight given to new sequences in the alignment at each iteration.

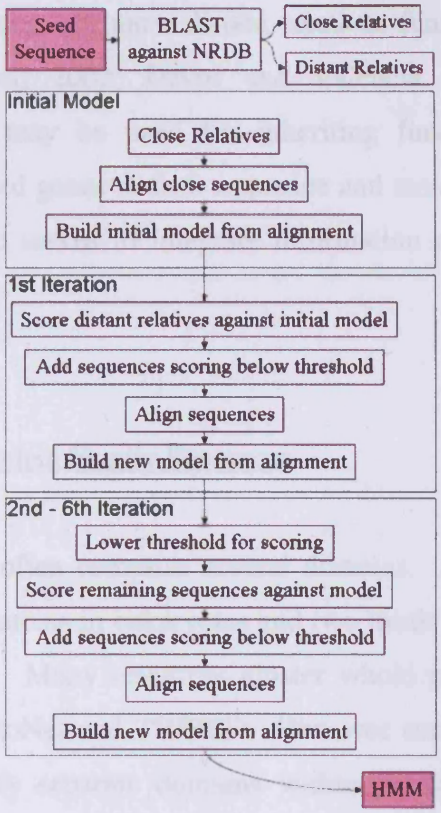


**Figure 2.2 Iterative Nature of the HMM Build Process.** *At each iteration an augmented alignment is produced that generates an improved model describing the sequence diversity of more distant relatives whilst maintaining the core sequence diversity of close relatives from the initial iteration.*

### 2.1.5 Protein Family Resources

Protein sequence mutations such as point mutations or larger insertions and deletions during evolution have given rise to families of proteins within which relatives share a common evolutionary ancestor, but may have diverse protein sequence and structure, and subsequent modification of protein function, although some protein families, notably the globins retain a common biological function despite high sequence diversity.

The term protein family has been in use since the 1960s *e.g.* (Dayhoff (ed) 1965-1978, Atlas of Protein Sequence and Structure) and definitions vary. The term is used here simply to refer to groups of proteins related by common ancestry, including close and distant relatives. Close relatives in the same protein family often possess similar or

identical biological functions. In more distant relatives functions may have changed (Todd *et al.*, 2001; Rost, 2002; Devos and Valencia, 2000). Protein family bioinformatics resources may be used for inheriting functional information from experimentally characterised genes to their sequence and structural relatives. Grouping proteins into families also serves to integrate information on cellular and molecular function.

#### 2.1.5.1 Sequence Based Protein Family Resources

Protein sequences often comprise several domains. Apic *et al.* (2001) predict that about four fifths of proteins in eukaryotes and two thirds of proteins in prokaryotes are multidomain proteins. Many resources cluster whole protein chains into protein families, for example ProtoNet and TRIBES. However some resources, for example Pfam and ADDA, identify separate domains within proteins and construct protein domain families. A single protein may consist of several protein domains that belong to different domain families. Hence protein domain family classifications are a useful tool in determining the evolutionary relationships between proteins, especially within a genomic context. Table 2.0 summarises some major sequence family databases.

**Table 2.0 Summary of Example Sequence Family Databases**

| Resource or Clustering Method | Reference | Source(s) | Families (06/04) | Method |
|---|---|---|---|---|
| ADDA | Heger & Holm | SWISSPROT, TrEMBL, PIR, PDB, WORMPEP, ENSEMBL | 34,000 families (plus 60,000 singleton) | BLAST |
| CHOP | Liu & Rost | 62 complete genomes | 63,300 clusters (plus 118,108 singletons) | PSI-BLAST |
| COG/KOG | Tatusov *et al.* | 66 unicelluar and 7 eukaryotic complete genomes | 4873 COG, 4852 KOG | Bidirectional best hit by BLAST |
| DIVCLUS | Park and Teichmann | 6 genomes, 12013 sequences | 13076 duplication modules in 1622 clusters | Multiple sequence comparison methods |
| InterPro | Mulder *et al.* | UniProt, PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR SuperFamily, SUPERFAMILY, Gene3D | 11,007 entries (2573 domains, 8166 families) | Multiple methods (HMM, PSI-BLAST, Regular Expression) |
| iProClass | Huang *et al.* | PIR, SWISSPROT, TrEMBL, Pfam, BLOCKS, PRINTS, ProSite, PDB, COG | 36,000 PIR superfamilies, 100,000 families | N/A |
| Pfam | Bateman *et al.* | SWISSPROT, TrEMBL | 7459 families | HMM |
| PRINTS | Attwood *et al.* | SWISSPROT, TrEMBL | 1800 entries 10,931 motifs | Iterative motif searches |

| | | | | |
|---|---|---|---|---|
| ProDom | Servant *et al.* | SWISSPROT, TrEMBL | 501,917 families (186,303 non-singleton) | PSI-BLAST |
| ProtoNet | Kaplan *et al.* | SWISSPROT, TrEMBL | User-defined | BLAST |
| SMART | Letunic *et al.* | Selected proteins | 667 domains | HMM |
| SwissPROT | Boeckmann *et al.* | Primary database | 153,871 proteins | N/A |
| SYSTERS | Meinel *et al.* | SWISSPROT, TrEMBL, ENSEMBL (complete genomes), the Arabidopsis Information Resource, SGD and GeneDB | 158,153 disjoint clusters | BLAST |
| TIGRFAMs | Haft *et al.* | SWISSPROT, TrEMBL | 1976 families | HMM |
| TRIBES | Enright *et al.* | 83 Complete Genomes | 60,934 or 82,692 depending on granularity | TribeMCL clustering using BLAST |

2.1.5.2 Families of Whole Protein Sequences

The TRIBES database (Enright *et al.*, 2003) clusters proteins from 83 complete genomes into between 60,934 and 82,692 families depending upon the level of granularity of clustering that is chosen. TRIBES uses the TribeMCL (Enright *et al.*, 2002) clustering program, developed from the Markov cluster (MCL) algorithm (van Dongen, 2000) to cluster protein sequences. TribeMCL simulates flow in a similarity graph consisting of pairwise sequence similarities (BLAST E-value cut-off of 0.0001) of all proteins in the dataset and then assigns complete protein sequences into families based on the density and strength of links between them. This novel approach does not suffer greatly from the problems caused by multi-domain proteins, promiscuous

domains and fragmented proteins as outlined previously. The method makes no attempt to decompose the sequences into their component domains but rather produces clusters that correlate well with the overall domain architecture of the sequences.

ProtoNet developed by Linial and co-workers (Kaplan *et al.*, 2005), clusters SWISSPROT proteins in the UniProt database on the basis of sequence similarity. Proteins from the TrEMBL repository are later added into these initial protein clusters. The ProtoNet protocol can produce protein family clusters from three different clustering methods: harmonic, geometric and arithmetic. These different clustering methods vary by putting more emphasis on either strong sequence similarities between cluster members (harmonic > geometric > arithmetic), or conversely on weak similarity between cluster members (arithmetic > geometric > harmonic) when merging clusters (Sasson *et al.*, 2003).

The PRINTS database (Attwood *et al.*, 2003) is a collection of protein 'fingerprints': regular expressions describing conserved sequence motifs used to characterise a protein family. These motifs are generated via multiple protein sequence alignments by identifying regions of local sequence conservation. They can subsequently be used to scan a larger sequence set (SWISSPROT and TrEMBL, Boeckmann *et al.*, 2003) to recruit new family members. The majority of families are defined by multiple motifs and all must be present for a relative to be added to the group.

COG and KOG (Tatusov *et al.*, 2004) are databases of clusters of orthologous groups of proteins, defined by bi-directional best hitting groups of three or more proteins in complete genomes (described previously in section 1.3.4).

The SYSTERS (Meinel *et al.*, 2005) database uses graph-based methods to generate protein families of varying granularity.

The number of families identified by those resources performing automated clustering of large sequence repositories varies from 65,000 to 186,000 depending on the clustering philosophy. Kunin and co-workers recently revealed that each newly sequenced genome leads to an increase in the total number of protein families characterised (Kunin *et al.*, 2003). A proportion of protein sequences (between 10 and

25%), in every genome, are singletons or belong to families not present in any other sequenced genome. This may reflect limitations in the current sequence based homologue detection algorithms; or alternatively these may be genuinely novel families that have arisen following speciation. These organism-specific families may be important for expanding the functional repertoire and phenotype of the organism, perhaps by providing new biological processes or changes in gene/protein regulation.

2.1.5.3 Families of Protein Domain Sequences

A number of resources exist which automatically cluster protein sequences from the completed genomes or from the large sequence repositories (e.g. GenBank or SWISSPROT-TrEMBL) into putative domain families. The ProDom resource (Servant *et al.* 2002) contains protein sequence families derived from sequences in UniProt and TrEMBL. These protein sequences are chopped into protein domains using an iterative PSI-BLAST domain boundary prediction program.

DIVCLUS (Park and Teichmann, 1998) is part of the Genome Analysis and Protein Family Maker software package that identifies homologous domains in single and multi-domain proteins. DIVCLUS uses an iterative checking process that compares pairs of aligned sequences and separates single linkage clusters into duplication module families according to sequence similarity and overlap criteria to produce clusters containing homologous proteins.

Pfam (Bateman *et al.* 2004) is a highly comprehensive resource providing a high quality set of Hidden Markov Model profiles for protein domain families. Families are built from ProDom identified clusters. These families are defined in Pfam using multiple sequence alignments and HMMs, the largest domain families are built first. Pfam consists of two parts, the first is the curated part of Pfam (Pfam-A), the second is an automatically generated supplement called Pfam-B.

TIGRFAM (Haft *et al.*, 2003) protein families are built in a similar fashion to Pfam but also contain whole protein chains.

Holm and co-workers recently developed the ADDA algorithm to cluster sequences into domain families by exploiting the principal of domain recurrence in different protein sequences (Heger and Holm, 2003). A related algorithm, CHOP (Liu and Rost, 2004), assigns domain boundaries by BLAST sequence comparison and then clusters the subsequent domain-like fragments into sequence families using the CLUP (Liu and Rost, 2004) clustering method (these methods are discussed in section 4.1).

SMART (Letunic *et al.*, 2002) (Simple Modular Architecture Research Tool) describes over 600 domain families, which have been selected with a particular emphasis on mobile eukaryotic domains and as such are widely found among nuclear, signalling and extracellular proteins. SMART domain families are defined by hand curated multiple sequence alignments. An HMM library of these domain families allows fast sequence annotation with SMART domains. SMART domain families are annotated with function, sub-cellular localisation, phylogenetic distribution and tertiary structure.

The InterPro database (Mulder *et al.*, 2005) is an important recent development since it integrates major protein family classifications and provides regular mappings from these resources to primary sequences in the UniProt protein sequence database. Databases in the InterPro collaboration include UniProt, PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR SuperFamily, SUPERFAMILY and more recently Gene3D.

#### 2.1.5.4 Structure Based Protein Family Resources

There are two major protein structure classifications, both of which require a varying degree of manual intervention, CATH and SCOP, which classify protein domains of known structure into evolutionary superfamilies. Each superfamily is classified in a hierarchy corresponding to Class (proportion of alpha helices and beta strands in the structure) and Fold Group (structures sharing significant global secondary structural element composition and connectivity). Architecture, (overall shape of structures, i.e. orientation of the secondary structures, for example layered sandwich or barrel like), is an additional hierarchical level in the CATH classification. Domains adopting similar Class, Fold Group and Architecture can be further clustered into

Homologous Superfamilies according to further evidence of an evolutionary relationship, for example shared functional characteristics and/or sequence motifs. Table 2.1 summarises both manual and automated structure based resources.

**Table 2.1 Protein Structure Family Resources**

| Database | Coverage (07/04) | Structural Comparison Method | Description |
|---|---|---|---|
| CAMPASS | 7580 domains<br>1409 superfamilies | COMPARER<br>SEA | Structure-based alignment of SCOP superfamilies |
| CATH | 67,054 domains<br>1572 superfamilies,<br>907 folds | SSAP<br>GRATH | Automatic structural and sequence comparison with manual validation of superfamily alignments and domain boundaries |
| CE | All chains in PDB | CE | Fully automatic, nearest neighbours |
| DALI Domain Dictionary | 1,062 superfamilies | DALI | Fully automatic classification using PUU, DALI algorithms |
| DHS | 1459 superfamilies | SSAP<br>CORA | Fully automatic multiple structural alignments of close relatives in CATH |
| HOMSTRAD | 7500 domains<br>1032 superfamilies | COMPARER | Manual classification of close protein homologues |
| MMDB | 28,000 structures,<br>87,000 domains | VAST | Fully automatic, nearest neighbours |
| SCOP | 65,122 domains<br>1325 superfamilies,<br>805 folds | Manual | Manual classification |

The SCOP database (Andreeva *et al.*, 2004) uses almost entirely manual validation for recognising structural similarities between distantly related protein structures to generate evolutionary superfamilies, resulting in a very high quality resource. In the CATH database, (Pearl *et al.*, 2005), a combination of manual and automated approaches is used. Whilst structure comparison methods (SSAP, Orengo and Taylor, 1996; CORA, Orengo, 1999; GRATH, Harrison *et al.*, 2002) have been developed to recognise structural relatives, evolutionary relationships are only assigned following manual assessment of all available data. Table 2.1 above shows that SCOP and CATH recognise around 805 - 907 fold groups and around 1325 - 1572 superfamilies in the current set of protein structures.

In contrast to the manually curated SCOP and CATH classifications, the DALI domain database (Dietmann *et al.*, 2001) is produced by a completely automated classification protocol. Domain boundaries are recognised using the PUU algorithm (Holm and Sander, 1994) and domains are assigned to fold groups and superfamilies using the robust DALI structure comparison algorithm (Holm and Sander, 1994).

The NCBI Molecular Modelling DataBase (MMDB, Chen *et al.*, 2003) classifies all non-theoretical PDB structures. Fully automatic classification in the MMDB is achieved using all against all structural comparison by secondary structure element superposition (VAST, Gibrat *et al.*, 1996). MMDB structures are linked to other NCBI databases containing sequences, taxonomy, literature references and both sequence and structure relatives.

CE (Shindyalov and Bourne, 1998) classifies PDB structures by structural comparisons using the CE algorithm that compares alpha carbon atom positions in the peptide chain to identify aligned fragment pairs. The optimal alignment of aligned fragment pairs is calculated by minimising RMSD.

Functional annotation of structural domains in CATH is achieved in the Dictionary of Homologous Superfamilies (DHS, Bray *et al.*, 2000). The DHS contains multiple structural alignments of all the known domain structures in each CATH homologous superfamily. Functional annotation is provided by BLASTing domain sequences against the UniProt database to identify 95% sequence identity relatives. Functional annotations from KEGG, COG, GO and EC numbers are then inherited from

the annotated UniProt sequence relative to the CATH domain. The current release of the DHS contains 1459 CATH homologous superfamily domains which have been annotated with 495,611 UniProt sequence relatives.

The HOMSTRAD and CAMPASS databases, constructed by Blundell and co-workers (Mizuguchi *et al.* 1998; Sowdhamini *et al.* 1998), are not hierarchical but focus on using SCOP, PFAM and other resources to cluster together families of evolutionary relatives. HOMSTRAD (HOMologous STRucture Alignment Database) groups proteins into families on the basis of sequence and structural similarity. HOMSTRAD combines SCOP, Pfam, PROSITE and SMART classifications with PSI-BLAST sequence similarities and sequence-structure profiles to define protein families. Currently the PDB is grouped into 1032 families representing 3454 structures. Each family is represented by manually curated structure-based alignments. The CAMPASS database groups more distant structural homologues than HOMSTRAD by using the structural comparison algorithms COMPARER and SEA to generate multiple alignments of SCOP superfamilies. Both these resources contain validated multiple structural alignments for families and superfamilies that can be used to identify further relatives using derived substitution matrices or conserved structural features.

### 2.1.6 Structural Annotation of Genomes

More than two thousand high throughput complete genome sequencing initiatives have produced over a million protein sequences. These protein sequences require annotation. Such annotation must be assigned using methods that are fast and reliable and the large amount of resulting data must be processed and stored efficiently. The assignment of structure to newly sequenced proteins using fast automated assignment methods, notably PSI-BLAST and the HMM's described previously, has resulted in development of genome annotation databases that contain sequence-based and structure-based annotation of complete genomes. There are four main genome annotation databases that contain structurally annotated genomes. These are SUPERFAMILY, the Genomic Threading Database, 3D-GENOMICS database and Gene3D. The processes used for constructing each database are described below.

The SUPERFAMILY database (Madera *et al.*, 2004) uses SAMT generated profile hidden Markov models to predict SCOP structural domains in 220 complete genomes. The Genomic Threading Database (McGuffin *et al.*, 2004) applies the GenThreader algorithm (described previously, see section 1.2.2.2) to assign structural domains to 218 complete genomes. The 3D-GENOMICS database (Fleming *et al.*, 2004), as of January 2005, contains 173 complete genomes. These genomes are annotated with SCOP structural domains (using PSI-BLAST derived IMPALA profiles) and Pfam sequence domains (using HMMER) as well as PROSITE and COG functional annotation. In addition, protein sequences are annotated with coiled-coil, low complexity, signal peptide, secondary structure, repeated regions and transmembrane regions.

Gene3D was originally set up by Buchan *et al.* in 2002 (Buchan *et al.*, 2002; 2003) and contained 66 complete genomes (53 bacteria, 11 archaea and 2 eukaryota) and consisted of domain assignments from CATH (version 2.4) and Pfam (release 6.2) mapped onto protein sequences using PSI-BLAST derived IMPALA profiles.

The recent increase in the number of completed genomes, especially those of eukaryotes, necessitated further development of Gene3D to increase the number of genomes and to provide protein family information. This chapter describes development of a novel protocol (PFscape) for identification of both protein and domain families in complete genomes and the implementation of this protocol in building release 3 of Gene3D.

## 2.2 Objectives

This chapter reviews methods for clustering sequences into families of relatives using sequence and structure-based methods and the use of these clusters to annotate complete genome sequences. The design and use of a novel protocol called PFscape, for assigning protein family, domain annotation and functional annotation to complete genomes is described in the construction of Gene3D, a resource containing protein sequences from 120 complete genomes, clustered into protein families and annotated with structural and sequence domain family information, together with metabolic pathway and functional data from GO, KEGG and COG. Finally, the user interface and web services are briefly outlined.

## 2.3 Results

Gene3D was completely re-structured to include information on both protein families and domain families. Protein families are themselves sub-divided into sequence identity clusters of more closely related protein sequences. Within each protein family, sequence and structural domain assignments to each protein sequence reveal domain context and assign domain architectures to proteins. Gene3D differs from other structural genomics resources (described previously, see section 2.1.6) in that, in addition to domain assignments, proteins are clustered into protein families. The advantage of this clustering approach is that within a protein family, domain and functional assignments to annotated protein members can be inherited to un-annotated protein members. The use of subclusters within each protein family permits such inheritance to be restricted to protein relatives with varying degrees of sequence identity.

### 2.3.1 Genome Sources in Gene3D

Gene3D version 3.0 contains 120 complete genomes (90 bacteria, 14 archaea and 16 eukaryota) comprising 854,897 protein sequences. The majority of genomes were downloaded from the NCBI. However, seven eukaryotic genomes (*Takifugu rubripes*, *Arabidopsis thaliana*, *Homo sapiens*, *Drosophila melanogaster*, *Anopholes gambiae*, *Mus musculus*, and *Rattus norvegicus*) were downloaded from ENSEMBL, where the sequence collections were more recent. Subsequent version updates in Gene3D have added over 100 additional genomes.

### 2.3.2 Domain Sources in Gene3D

Each protein is annotated separately with CATH (release 2.5) and Pfam (release 10) protein domains. These two domain annotation schema can be combined into a Domain Architecture for each protein, where gaps in the CATH domain assignment are filled with non-overlapping Pfam domain assignments. Remaining unassigned regions of 50 residues or more are denoted as NewFam regions. CATH and Pfam assignments

are achieved using an HMM library containing 4036 models representing 1467 CATH structural domains and 6190 models representing 6190 Pfam sequence domains. Subsequent version updates in Gene3D have added CATH release 2.6 and Pfam release 17 domain assignments.

### 2.3.3 Family Clusters in Gene3D

Protein sequences in Gene3D are clustered into protein families using TribeMCL, a clustering program written by Enright *et al.* (Enright *et al.*, 2002). Each protein family is further clustered into sequence identity subclusters of 35, 60, 95 and 100 percent sequence identity using Homolseqs, in-house single linkage clustering software (Orengo *et al.*, 1997).

Domain assignments in Gene3D are clustered into domain families on the basis of their CATH or Pfam domain family classification. Each domain family is further clustered into sequence identity subclusters of 30, 35, 40, 50, 60, 70, 80, 90, 95 and 100 percent sequence identity using TCluster, in-house directed multi-linkage clustering software written by Tony Lewis.

### 2.3.4 Database Tables in Gene3D

In order to facilitate complex queries a Gene3D relational database was designed to store both domain assignment data and protein family data. The core database consists of four MySQL (open source database, www.mysql.com) tables, illustrated in figure 2.3. The central PROTEIN table stores data for all the 854,897 protein sequences in the database. This table links to DOMAIN ASSIGNMENT and PROTEIN FAMILY tables to show domain assignments and protein family classifications respectively for each protein. Lastly, the ORGANISM table designates the genomic context of each protein in the PROTEINS table.

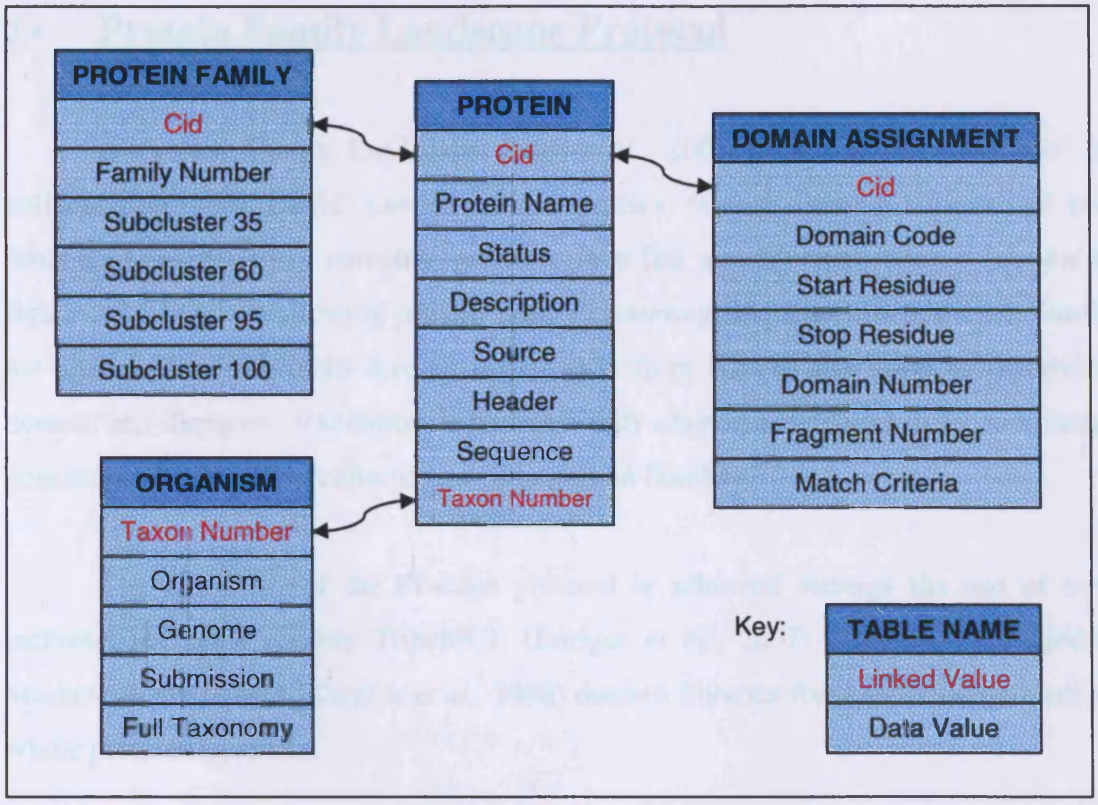


**Figure 2.3 Gene3D Database Table Structure.** *Note that red linked values allow data in different tables to be connected (arrows).*

Table links occur between Cid and Taxon Number data fields. Organisms in Gene3D are designated using NCBI Taxonomy (Benson *et al.*, 2000) taxon numbers. The taxon number field provides a unique identifier for each organism. Genomes downloaded from the NCBI (Wheeler *et al.*, 2005) were directly assigned a taxon number from the NCBI Taxonomy resource. Where genomes were downloaded from ENSEMBL (Hubbard *et al.*, 2005), the taxon number was added manually.

Proteins in Gene3D have many different formats of protein identifiers. To facilitate manipulation of the data an internal protein identifier was used. A Cid is an internal identifier that is kept within the Gene3D database. Each protein sequence in the database is assigned a unique 8-digit Cid (for example: 00000001). Identical proteins that occur in different organisms or in different locations within the same organism are distinguished by different Cids.

## 2.4 Protein Family Landscape Protocol

A Protein Family Landscape (Lee *et al.*, 2005) protocol was developed in collaboration with David Lee to assign protein family, domain annotation and functional annotation to complete genomes in a fast and efficient manner (shown in figure 2.4 below). Following protein family clustering, sequences within each family are annotated with protein domain information from CATH and Pfam to determine domain architectures. Validation of protein family clusters is undertaken by comparing consistency of domain architectures within protein families.

The efficiency of the PFscape protocol is achieved through the use of new technologies, most notably TribeMCL (Enright *et al.*, 2002) clustering and hidden Markov model (HMM, Karplus *et al.*, 1998) domain libraries for domain assignment to whole protein sequences.

PFscape protocol generation of Gene3D is only possible due to the recent acquisition of computers with very large amounts of memory and a 150 processor computer farm where tasks can be massively parallelised. The PFscape protocol has three stages: protein family clustering, protein domain assignment and functional annotation. These are described below.

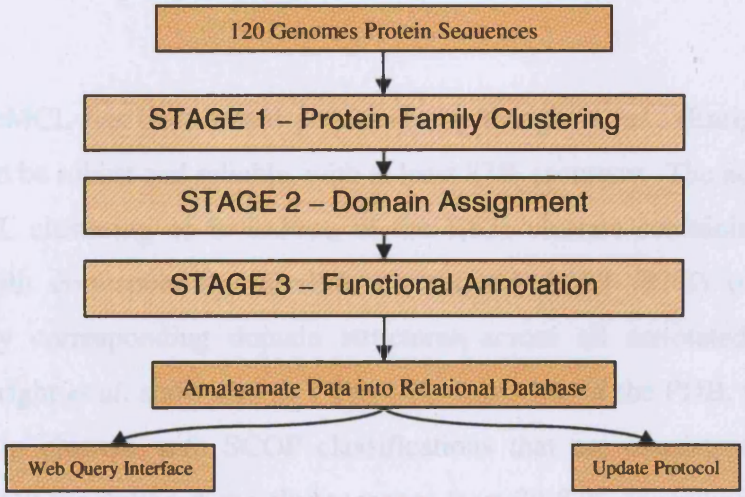


**Figure 2.4 PFscape Protocol Structure.** *Note that protein clustering and domain assignment are simultaneous independent processes.*

### 2.4.1 Stage 1: Protein Family Clustering

TribeMCL (Enright *et al.*, 2002) was chosen for protein family clustering as this method is fast, automated, and successfully overcomes the clustering problems of protein domains, fragmented peptides and sequence similarity errors (as discussed previously). TribeMCL is based upon the Markov clustering algorithm (van Dongen, 2000) and represents proteins as nodes in a graph; connections between these nodes represent sequence similarity between proteins. A matrix consisting of sequence similarities in the graph, transformed into probabilities associated with transitions from one protein to another within the graph, is passed through iterative cycles of matrix multiplication and matrix inflation to simulate random walks on the graph. Protein clusters in the graph can be identified since random walks on the graph are less likely to go between clusters than remain within a single cluster, since within protein clusters there is a higher probability of transitioning between cluster members than between members of different clusters. Matrix multiplication (i.e. matric squaring) computes longer random walks on the graph and associates new probabilities with all pairs of nodes in the graph, thus acting to dissipate clusters. Matrix inflation augments the probabilities of intra-cluster walks and diminishes the probabilities of inter-cluster walks, eliminating connections between clusters. Iterative rounds of expansion and inflation act to separate the graph into clusters. The granularity of these clusters can be altered by changing the matrix inflation parameter, to produce tighter or broader clusters.

TribeMCL has been tested previously by Enright *et al.*, (Enright *et al.*, 2002) and shown to be robust and reliable, with at least 87% accuracy. The authors report that in TribeMCL clustering of SwissProt, of the 1,821 clusters containing four or more members with corresponding InterPro annotations, 1,583 (87%) of these clusters contain fully corresponding domain structures across all annotated members. In addition, Enright *et al.* show that in TribeMCL clustering of the PDB, the total number of proteins in clusters with SCOP classifications that are consistent with the most common SCOP annotation in the cluster ranges from 79-87% depending on the inflation value used.

TribeMCL can cluster proteins at granularity levels between 1 (broadest clusters) to 3 (tightest clusters). Protein family clusters in Gene3D should contain

members where all proteins have the same domain architecture. In order to further validate TribeMCL, structural data was used as remote homologues are more reliably detected using structural data. The optimal granularity level for TribeMCL clustering was determined by validating clusters made at all three granularity levels against a dataset of structurally characterised proteins that had been manually validated in the CATH database.

#### 2.4.1.1 Benchmarking TribeMCL using Structural Data

Multi-domain protein structures from the PDB (having their individual domains previously classified in CATH) were clustered using TribeMCL at three levels of granularity. The resulting clusters were assessed by comparing the total percentage of proteins that contained the most common domain architecture (same CATH domain assignments in the same order) in each cluster. As can be seen in figure 2.5, a granularity level of 3 produces clusters where 94.8% of all proteins contain the most common domain architecture CATH classification in their cluster. Clustering at granularity level 3 is therefore the most appropriate to cluster proteins in Gene3D to obtain protein families with consistent domain architectures.

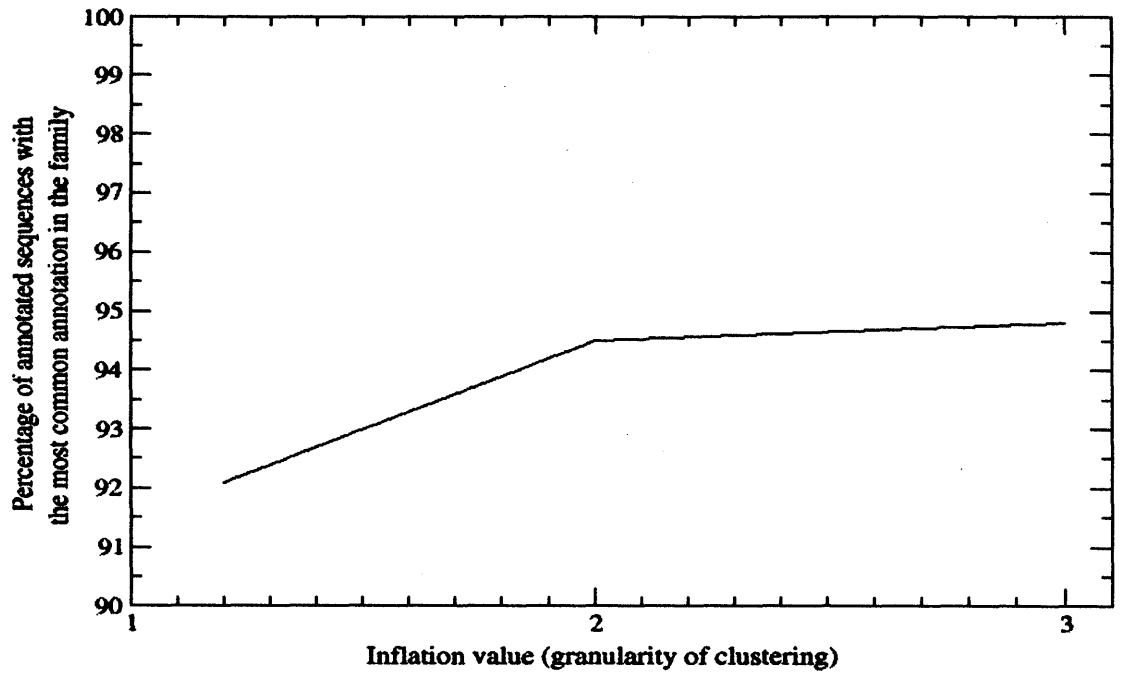


**Figure 2.5 TribeMCL Granularity Benchmarking.** *TribeMCL clustering of multidomain proteins in the PDB performed at different granularity levels. At granularity level 3, the broadest clustering level, 94.8% of all the proteins in the TribeMCL clusters contain the most common domain architecture in their cluster.*

Ideally the protein families identified by the PFscape protocol should consist of unique domain architectures, indicating that all identifiable evolutionary relatives in the PDB are present in the protein family. Half of all domain architectures are found to occur in a single cluster, see figure 2.6 below. This demonstrates that half of all domain architectures identified in the PDB are unique to a single cluster and are not found to occur in any other cluster. This effect is seen at all three inflation values, indicating that these clusters are quite robust and are formed early in the TribeMCL clustering process by distinct, closely related groups of proteins.

Figure 2.6 shows that TribeMCL clustering is quite conservative, 50% of domain architectures occur in two or more clusters and that therefore half of protein family clusters may need to be merged subsequently. However, this was considered preferable to protein families containing inconsistent domain architectures. Information on domain architecture from CATH, Pfam and NewFam domain family assignments will be exploited to merge protein families at some stage in the future (see section 3.3.4.2 and 6.2).

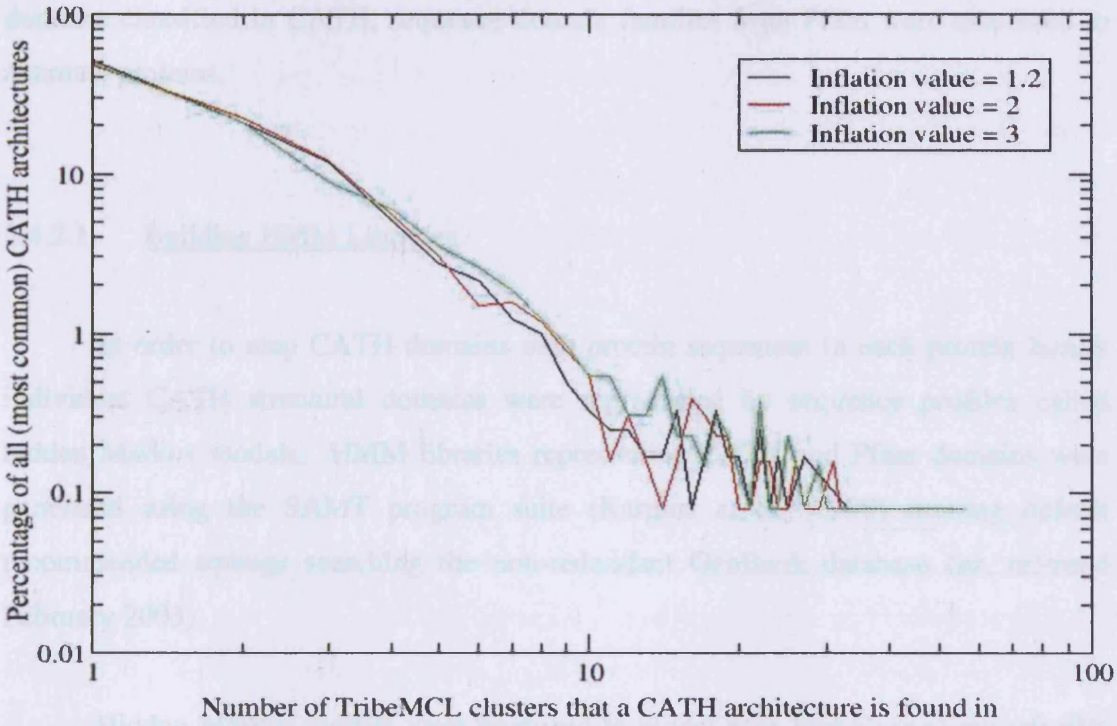


**Figure 2.6 TribeMCL Granularity Benchmarking.** *TribeMCL clustering of the Protein Data Bank performed at three different granularity levels. Note that half of CATH domain architectures are found in a single cluster irrespective of cluster granularity.*

### 2.4.2 Stage 2: Domain Assignment

Mapping protein domains onto protein sequences allows the conservative protein families produced from Stage 1 to be validated and confirm that protein families are, as the benchmarking would indicate, families containing evolutionary related proteins sharing common domain architecture. Additionally, domain assignment permits related protein family clusters to be more safely merged, where appropriate.

Protein domain regions are often more reliably identified by using structural data as the structures of proteins are much more highly conserved during evolution than the sequences of proteins. Whilst a number of comprehensive domain structure classifications exist, the CATH protein structure classification was chosen as the primary protein domain assignment classification in Gene3D. In addition to structural

domains classified in CATH, sequence domain families from Pfam were also used to annotate proteins.

#### 2.4.2.1 Building HMM Libraries

In order to map CATH domains onto protein sequences in each protein family individual CATH structural domains were represented by sequence profiles called hidden Markov models. HMM libraries representing CATH and Pfam domains were generated using the SAMT program suite (Karplus *et al.*, 1998) running default recommended settings searching the non-redundant GenBank database (nr, released February 2003).

Hidden Markov models were produced to model each homologous superfamily in CATH according to the schema shown in figure 2.7.

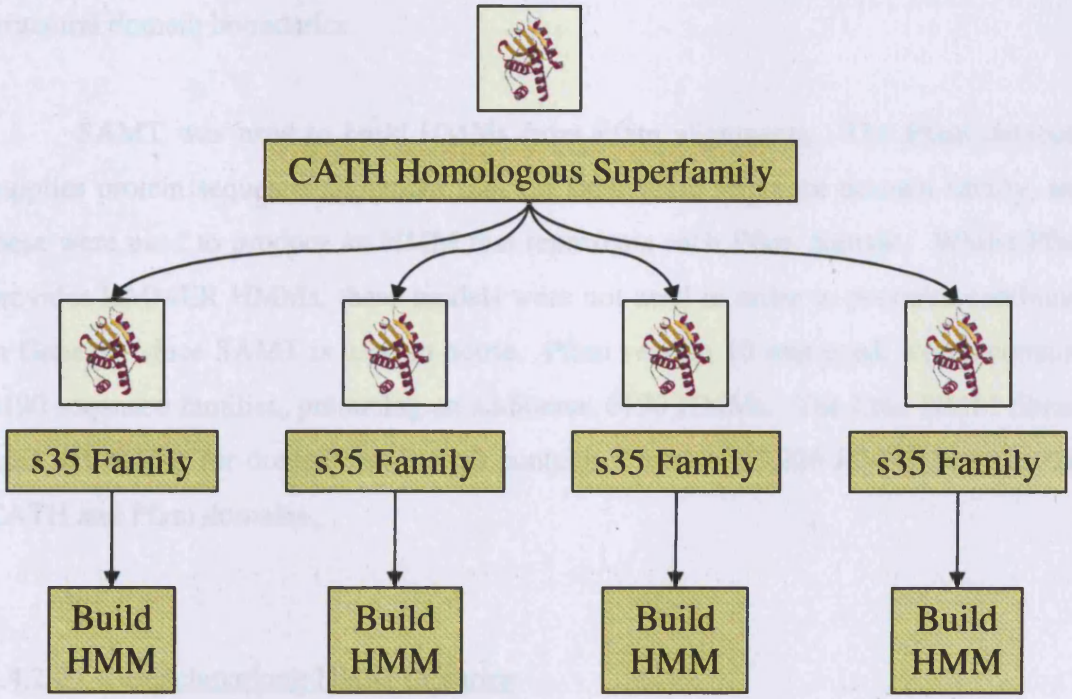


**Figure 2.7 HMM Representation of a CATH Homologous Superfamily.** *A representative domain sequence from each CATH s35 sequence family for every homologous superfamily in CATH is used to build an HMM. Multiple HMMs can therefore represent a single homologous superfamily.*

In order to capture all the protein sequence diversity present in a given CATH homologous superfamily, HMMs were built from representative sequences for each s35 family within a homologous superfamily (see section 2.1.5.4). A single homologous superfamily can therefore be represented using several HMMs, one for each s35 family. In total 4036 HMMs were produced representing 1467 homologous superfamilies in CATH.

Previous analyses have shown that on average, 35-45% of proteins within a genome contain structural domain assignments (Buchan *et al.*, 2003; Gough *et al.*, 2001). Therefore an additional source of protein domains for assignment is needed to increase genome coverage. Sequence/functional domains sometimes comprise just a fragment of a structural domain or in some cases correspond to multiple structural domains. There are several resources dedicated to sequence/functional domains in proteins which could be used for annotating genomes. Pfam provides a comprehensive and high quality set of profiles for protein domain families and now attempts, where structures are known, to ensure that sequence domain boundaries correspond with structural domain boundaries.

SAMT was used to build HMMs from Pfam alignments. The Pfam database supplies protein sequence alignment files for each Pfam sequence domain family, and these were used to produce an HMM that represents each Pfam domain. Whilst Pfam provides HMMER HMMs, these models were not used in order to preserve continuity in Gene3D, since SAMT is used in-house. Pfam version 10 was used, which contains 6190 sequence families, producing an additional 6190 HMMs. The final HMM library used in Gene3D for domain assignment contains a total of 10,226 HMMs representing CATH and Pfam domains.

#### 2.4.2.2 Benchmarking HMM Libraries

HMMs generated using the SAMT technology developed by Karplus *et al.*, are capable of identifying significantly more distant homologues than other profile methods, for example PSI-BLAST (Park *et al.*, 1998; Madera and Gough, 2002; see section 1.2.1.5). SAMT was benchmarked in-house by Sillitoe *et al.* (Sillitoe *et al.*, 2005). Figure 2.8

shows the accuracy of HMMs in detecting remote homologues in the CATH database at various error rates. This error rate was measured as the percentage of false positives.

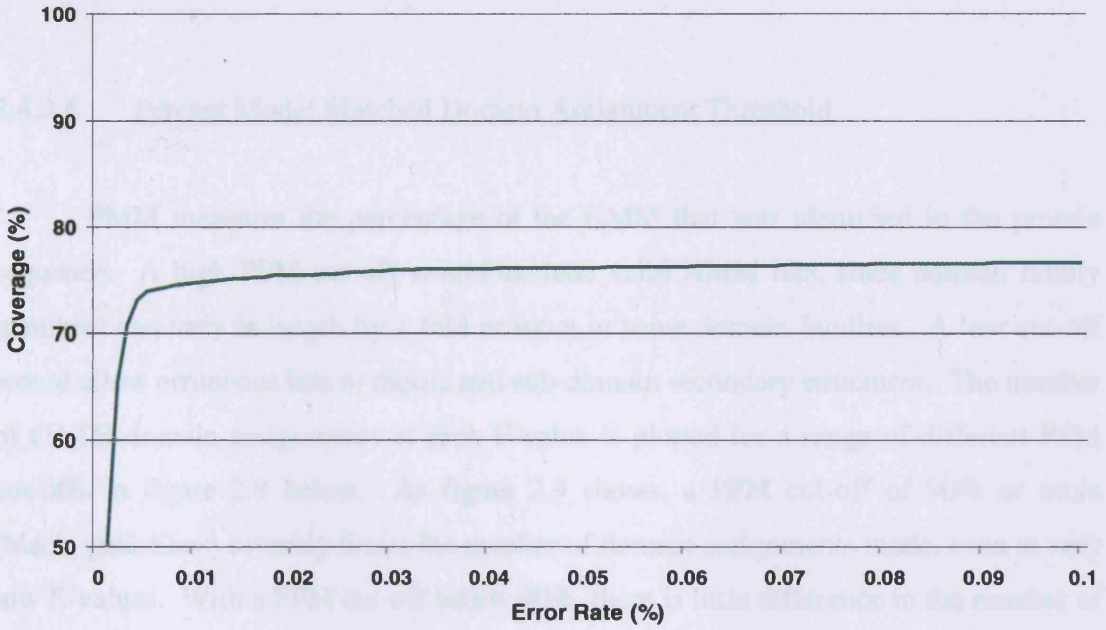


**Figure 2.8 HMM Coverage of CATH Homologous Superfamilies.** *The percentage of targets identified using HMM profiles. The dataset consists of CATH remote homologous superfamily members (< 35% sequence identity). Taken from Sillitoe et al, 2005.*

Scanning a remote homologue benchmark dataset against the CATH SAMT HMM model library identified 76% of homologues with an error rate of 0.1%.

### 2.4.2.3 Domain Assignment by DomainFinderII

The 854,897 protein sequences from 120 genomes in Gene3D were scanned against the HMM library using the SAMT program suite. SAMT output was filtered according to several criteria by DomainFinderII, an updated version of DomainFinder (Pearl *et al.*, 2002) in-house software written and updated by David Lee, and CATH and Pfam domain assignments were then made to protein sequences. DomainFinderII filters SAMT output using three criteria: (i) the percentage of the HMM model sequence that was identified in the protein sequence (Percent Model Matched (PMM)), (ii) the

percentage of the domain assignment that overlaps with other domain assignments in the protein sequence (Acceptable overlap (AO)) and finally (iii) the SAMT E-value for the domain assignment (E-value). These selection criteria are described below.

#### 2.4.2.4 Percent Model Matched Domain Assignment Threshold

PMM measures the percentage of the HMM that was identified in the protein sequence. A high PPM cut-off would exclude valid HMM hits, since domain family members can vary in length by 2 fold or more in some domain families. A low cut-off would allow erroneous hits to motifs and sub-domain secondary structures. The number of CATH domain assignments at each E-value is plotted for a range of different PPM cut-offs in figure 2.9 below. As figure 2.9 shows, a PPM cut-off of 90% or more (black, pink lines) severely limits the number of domain assignments made, even at very low E-values. With a PPM cut-off below 90%, there is little difference in the number of domain assignments made at E-values lower than 1.0e-10. Above E-values of 1.0e-10, the PPM cut-off has a marked effect on the number of assignments (see figure 2.9 (bottom)). Whilst a lower PPM cut-off allows many more domain assignments to be made (for example at an E-value of 1.0e-05, a PPM of 50% (orange line) allows 5000 more domain assignments to be made than a PPM of 80% (yellow line)), a PPM cut-off of 50% was chosen to ensure that at least half the domain defined by the HMM is identified in the sequence. This cut-off avoids domain families with members of variable length being penalised.

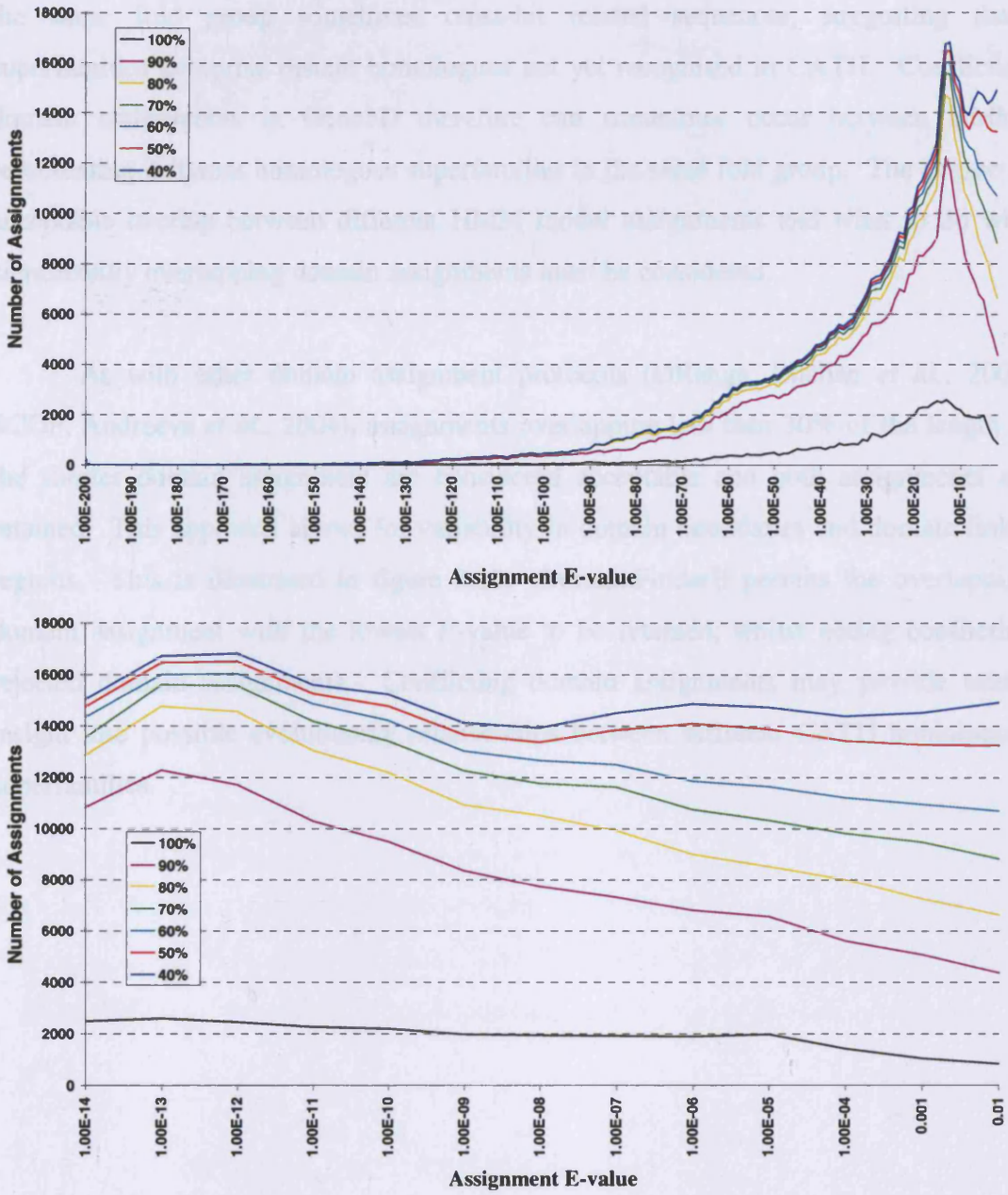


**Figure 2.9 Percent Model Matched Cut-off.** *The number of CATH domain assignments permitted by DomainFinderII at different percent model matched cut-offs is shown for all E-values (top) and expanded for E-values above 1.0e-14 (bottom).*

### 2.4.2.5 Acceptable Overlap Domain Assignment Threshold

Domain structures in the same CATH fold group with no evidence of an evolutionary relationship at the time of classification are placed into different homologous superfamilies. HMMs from different CATH homologous superfamilies in

the same fold group sometimes cross-hit related sequences, suggesting these superfamilies comprise distant homologues not yet recognised in CATH. Conflicting domain assignments in Gene3D therefore can sometimes occur between HMMs representing different homologous superfamilies in the same fold group. The degree of acceptable overlap between different HMM model assignments and what to do with significantly overlapping domain assignments must be considered.

As with other domain assignment protocols (DRange, Buchan *et al.*, 2002; SCOP, Andreeva *et al.*, 2004), assignments overlapping less than 30% of the length of the shorter domain assignment are considered acceptable and both assignments are retained. This approach allows for variability in domain boundaries and domain linker regions. This is illustrated in figure 2.10. DomainFinderII permits the overlapping domain assignment with the lowest E-value to be retained, whilst noting conflicting rejected domain assignments. Conflicting domain assignments may provide useful insight into possible evolutionary relationships between different CATH homologous superfamilies.

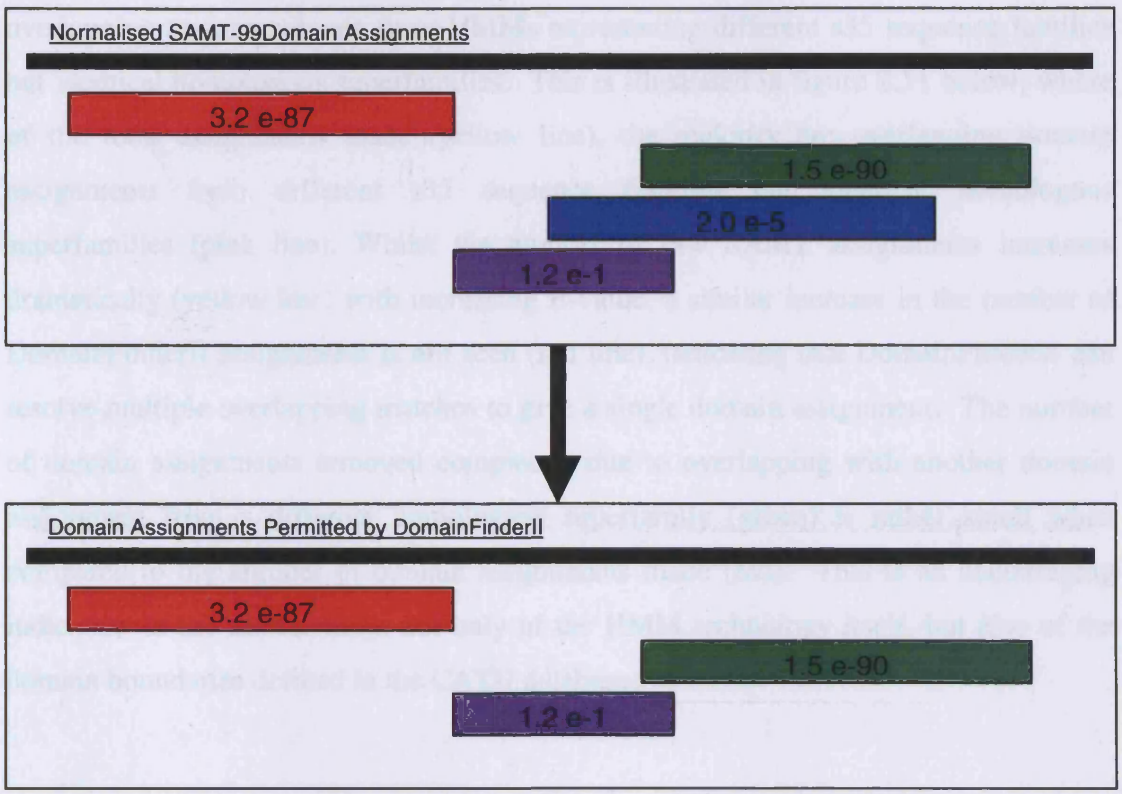


**Figure 2.10 Acceptable Overlap in DomainFinderII Domain Assignment.** *Normalised SAMT assignments (top box) are processed in order of lowest to highest E-value to produce DomainFinderII domain assignments (lower box). The non-overlapping red domain is permitted, whilst the blue domain is discarded due to unacceptable overlap with the green domain, leaving the purple domain acceptable.*

#### 2.4.2.6 E-value Domain Assignment Cut-Off

From benchmarking of the HMM library, (Sillitoe *et al.*, 2005), it was determined that an E-value cut-off of 0.01 was appropriate for the HMM library, producing accurate domain assignments with a 0.1% error rate.

#### 2.4.2.7 Resolving Multiple Overlapping Assignments

Normalised SAMT assignments comprise hits of HMMs representing Pfam domain families and CATH s35 sequence families. Some of the CATH s35 sequence families are members of the same CATH homologous superfamilies. These HMMs are likely to assign domains to the same protein region. The vast majority of significantly

overlapping assignments are from HMMs representing different s35 sequence families but identical homologous superfamilies. This is illustrated in figure 2.11 below, where of the total assignments made (yellow line), the majority are overlapping domain assignments from different s35 sequence families but identical homologous superfamilies (pink line). Whilst the number of raw SAMT assignments increases dramatically (yellow line) with increasing E-value, a similar increase in the number of DomainFinderII assignments is not seen (red line), indicating that DomainFinderII can resolve multiple overlapping matches to give a single domain assignment. The number of domain assignments removed completely due to overlapping with another domain assignment from a different homologous superfamily (green) is rather small when compared to the number of domain assignments made (red). This is an encouraging indication of the effectiveness not only of the HMM technology itself, but also of the domain boundaries defined in the CATH database.

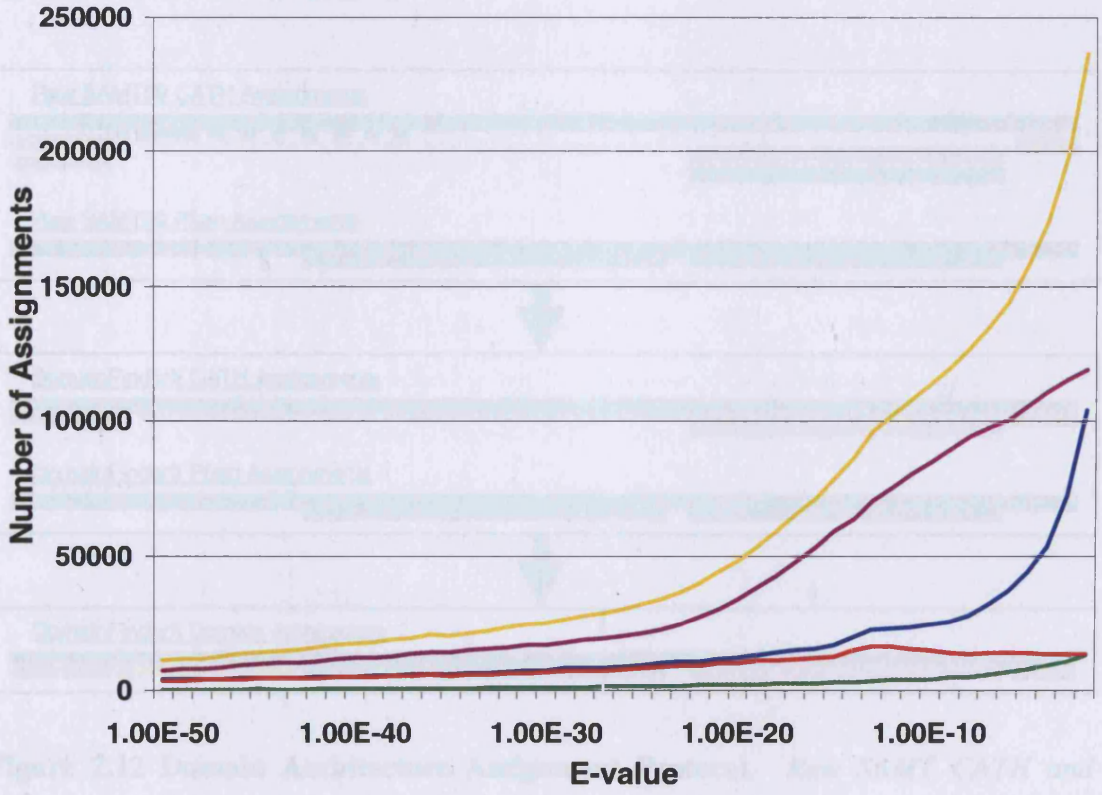


**Figure 2.11 DomainFinderII Effect.** *The number of CATH domain assignments: From the total number of SAMT assignments (yellow), assignments are discarded due to: overlapping hit to same homologous superfamily (pink), PPM cut-off (blue), overlapping hit to different homologous superfamily (green). The final assignment by DomainFinderII is shown in red.*

The domain assignment process described above was undertaken independently for CATH and Pfam HMM scan results using DomainFinderII. Domain assignments were then stored in the Gene3D database. The CATH and Pfam domain assignment tables were then used to construct domain architectures for protein sequences.

### 2.4.2.8 Domain Architectures

Domain architecture indicates the order of domains in the protein sequence to which they have been assigned. Priority is given to CATH domain assignments, Pfam domain assignments were added where they do not overlap significantly (less than 30% of the length of the shorter domain assignment). Unassigned regions were labelled as Newfam (unassigned putative domain) regions (discussed in more detail in section 3.3.2.1). Figure 2.12 illustrates the DomainFinderII domain assignment process and the identification of domain architectures.

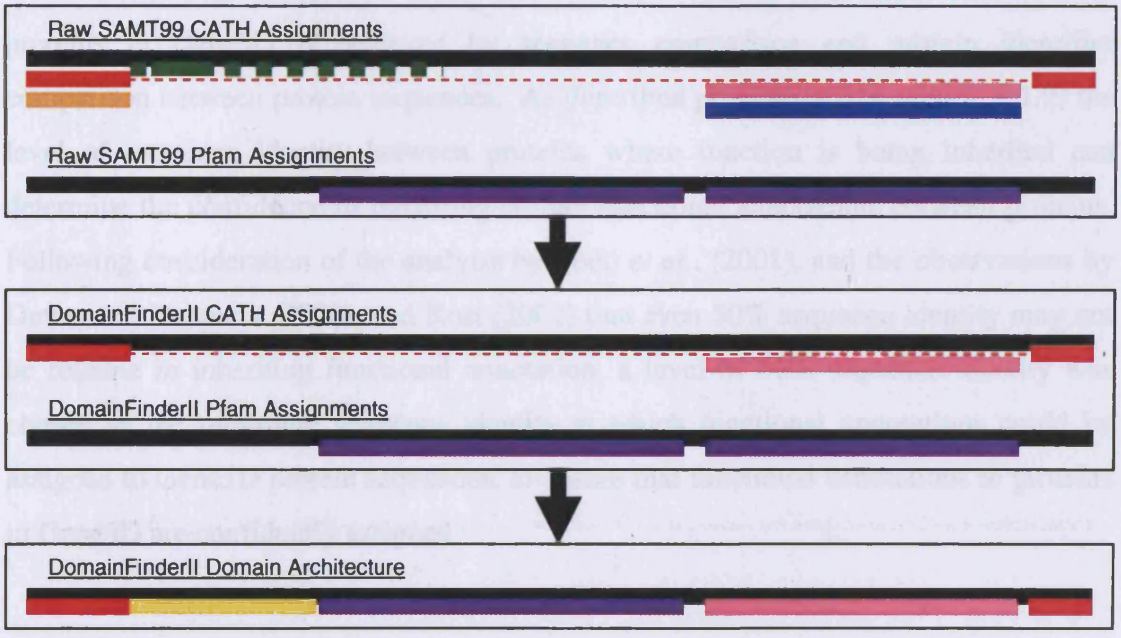


**Figure 2.12 Domain Architecture Assignment Protocol.** *Raw SAMT CATH and Pfam domain assignments to a protein sequence (black line) are filtered by DomainFinderII. Hits are discarded due to higher E-value same homologous family hit (pink), higher E-value different homologous superfamily hit (orange), and hit below PPM cut-off (green). Note that the red domain is discontinuous. Domain architecture assigned with priority to CATH domains (red, pink) before Pfam domains (purple). Finally, unassigned regions are labelled as Newfam domains (yellow).*

### 2.4.3 Stage 3: Functional Annotation

Gene3D is primarily intended for the study of protein and proteome evolution, by analysis of protein domain and protein family distributions across complete genomes. Functional assignment to protein sequences in Gene3D is sourced from resources that describe whole protein function, not only in molecular terms but also in biological process functions. Functional resources (described previously, see section 1.3) such as GO, KEGG and COG assign function via inheritance from sequence similar proteins whilst STRINGS and Affymetrix allow functional inheritance from proteins that may not be expected to have detectable sequence similarity. All five resources are used to assign functional information to protein sequences in Gene3D.

#### 2.4.3.1 Functional Assignment in Gene3D

Functional assignment from annotated proteins in functional resources to proteins in Gene3D is achieved by sequence comparison and protein identifier comparison between protein sequences. As described previously (see section 1.1.9) the level of sequence identity between proteins where function is being inherited can determine the confidence of inheriting correct functional annotations between proteins. Following consideration of the analysis by Todd *et al.*, (2001), and the observations by Devos and Valencia (2000), and Rost (2002) that even 50% sequence identity may not be reliable in inheriting functional annotation, a level of 60% sequence identity was chosen as the minimum sequence identity at which functional annotations could be assigned to Gene3D protein sequences, to ensure that functional annotations to proteins in Gene3D are confidently assigned.

#### 2.4.3.2 Function Assignment to Gene3D Proteins

STRINGS protein identifiers were mapped to KEGG and NCBI protein identifiers associated directly with proteins in Gene3D. GO, KEGG, COG and Affymetrix protein sequences were BLASTed against Gene3D protein sequences and functional assignments made where significant sequence similarities were identified (at least 80% of the longer sequence overlapped and there was 60, 95 or 100% sequence

identity). The level of sequence identity allows three different confidence levels to be assigned to the annotation.

A single Affymetrix microarray (RG_U34A, Rat Genome U34 Set) was used to give an indication of the percentage of microarray sequences that are represented in Gene3D. As figure 2.13 shows, almost three-quarters of the sequences represented on the microarray are found in Gene3D. This indicates that Gene3D is a resource that can be used to analyse microarray data, and also, that microarray data can be used to form functional associations between Gene3D proteins.

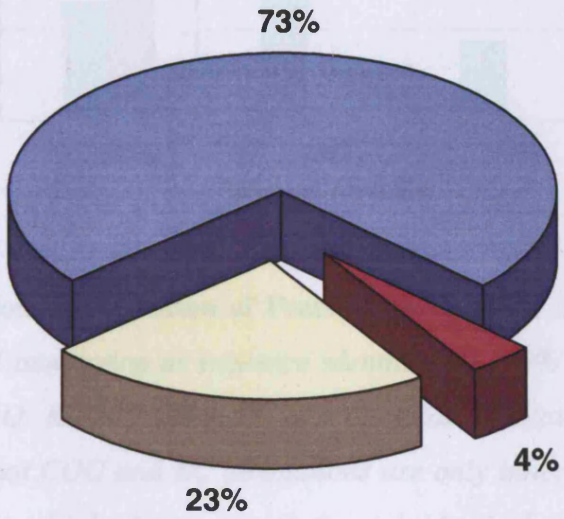


**Figure 2.13 Gene3D Coverage of Affymetrix Microarray.** *Percentage of genes represented on array that have a near identical sequence relative (BLAST E-value < 1.0 e-10, blue), a close sequence relative (BLAST E-value < 1.0 e-02, red) or no identifiable close sequence relative (BLAST E-value > 1.0 e-02, yellow) in Gene3D.*

Functional coverage of Gene3D is shown in figure 2.14. As the sequence identity cut-off at which functional annotation is inherited decreases from 100 percent (identical protein sequences) to 60 percent, there is a 9.1% (GO), 1.2% (KEGG) and 7.0% (Annotated) increase in the percentage of proteins in Gene3D that inherit functional annotation. These functional annotations were used to assign protein family names and functions and provide functional annotation for CATH domain families, described later.

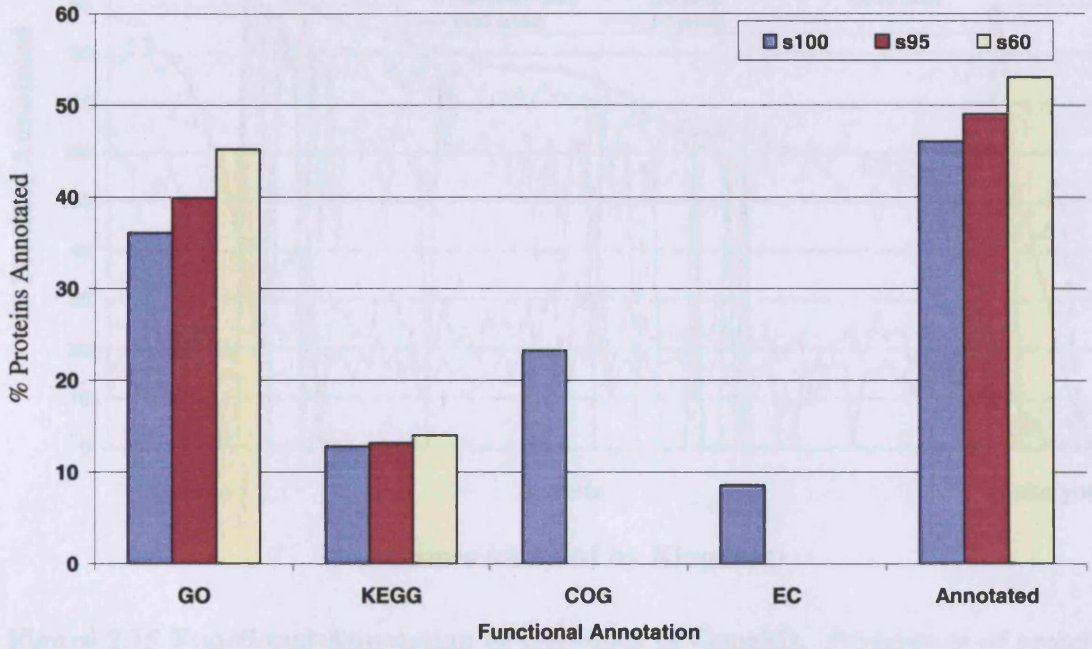


**Figure 2.14 Functional Annotation of Proteins in Gene3D.** *Percentage of proteins receiving functional annotation at sequence identities of 100% (blue), 95% (red) and 60% (yellow) for GO, KEGG, COG, EC and the total Annotated with any functional annotation. Note that COG and EC annotations are only inherited at 100% sequence identity since functional inheritance occurs through identical sequence identifiers (see section 2.4.3).*

### 2.4.3.3 Functional Coverage of Genomes

Whilst only 53% of total proteins in Gene3D have functional annotation, genome coverage of some individual genomes is more comprehensive. The percentage of proteins within a genome which have an associated function is shown in figure 2.15. Prokaryotic genomes are much better annotated than eukaryotic genomes. In addition, there is a much smaller increase in the number of additional proteins inheriting functional annotation at less stringent sequence identities in prokaryotic genomes than in eukaryotic genomes. This is due to a high degree of common prokaryotic sequences in Gene3D, GO and KEGG. All three resources use prokaryotic genome sources which are relatively stable, well curated sequence collections with many related organisms and are generally much better functionally characterised than eukaryotic genomes.

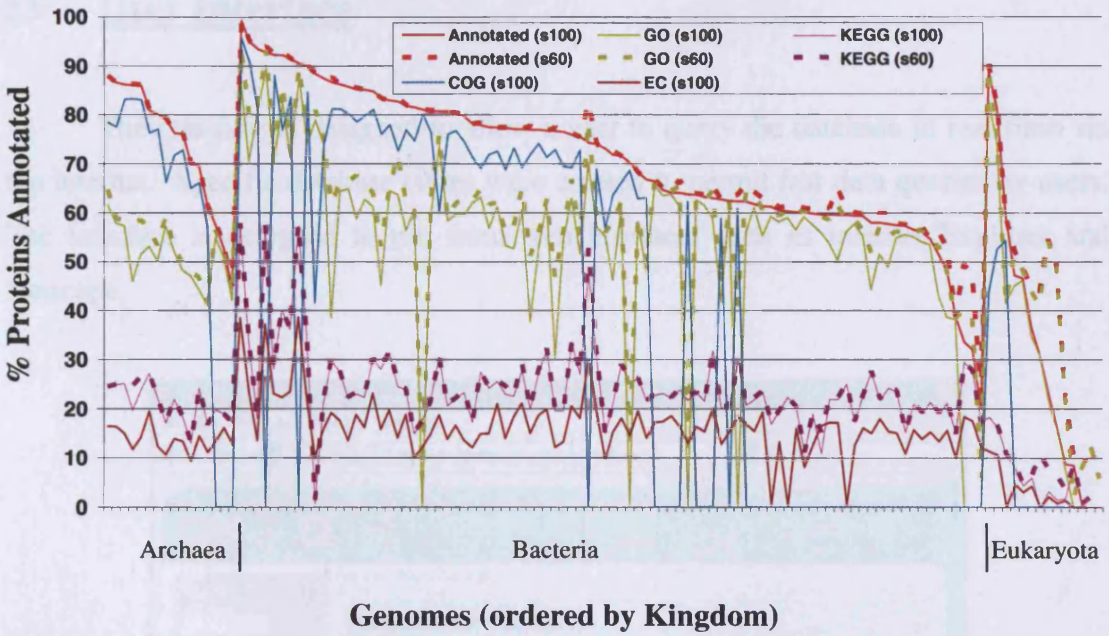


**Figure 2.15 Functional Annotation of Genomes in Gene3D.** *Percentage of proteins receiving functional annotation at sequence identities of 100% (line) and 60% (dotted line) for GO (green), KEGG (pink), COG (blue), EC (brown) and the total Annotated (red) with any functional annotation. Note that COG and EC annotations are only inherited at 100% sequence identity.*

Functional coverage of individual Kingdoms of life is shown in table 2.2. Higher eukaryotes are generally poorly annotated with the exception of GO annotations. Table 2.2 highlights the importance of functional annotation inheritance and shows that this has a marked effect on genome functional coverage in poorly annotated eukaryotic genomes where genome functional coverage increases by nearly 10% as functional annotations are inherited from 60% sequence identity relatives.

**Table 2.2 Functional Annotation by Kingdom in Gene3D.** *Functional annotation coverage at s100 and s60 inheritance.*

| Kingdom | %GO | %KEGG | %COG | %EC | % Annotated |
|---|---|---|---|---|---|
| Archaea (s100) | 50.9 | 20.9 | 59.8 | 13.7 | 73.8 |
| Bacteria (s100) | 54.5 | 24.7 | 40.5 | 16.6 | 70.7 |
| Eukaryota (s100) | 28.7 | 5.7 | 10.8 | 2.8 | 32.0 |
| Archaea (s60) | 53.6 | 21.4 | - | - | 73.9 |
| Bacteria (s60) | 59.7 | 25.1 | - | - | 71.8 |
| Eukaryota (s60) | 39.1 | 7.8 | - | - | 41.8 |

## 2.5 User Interface

The interface is designed to allow a user to query the database in real time via the internet. Specific database tables were created to permit fast data queries by users. The interface is designed to run from web browsers such as Internet Explorer and Netscape.

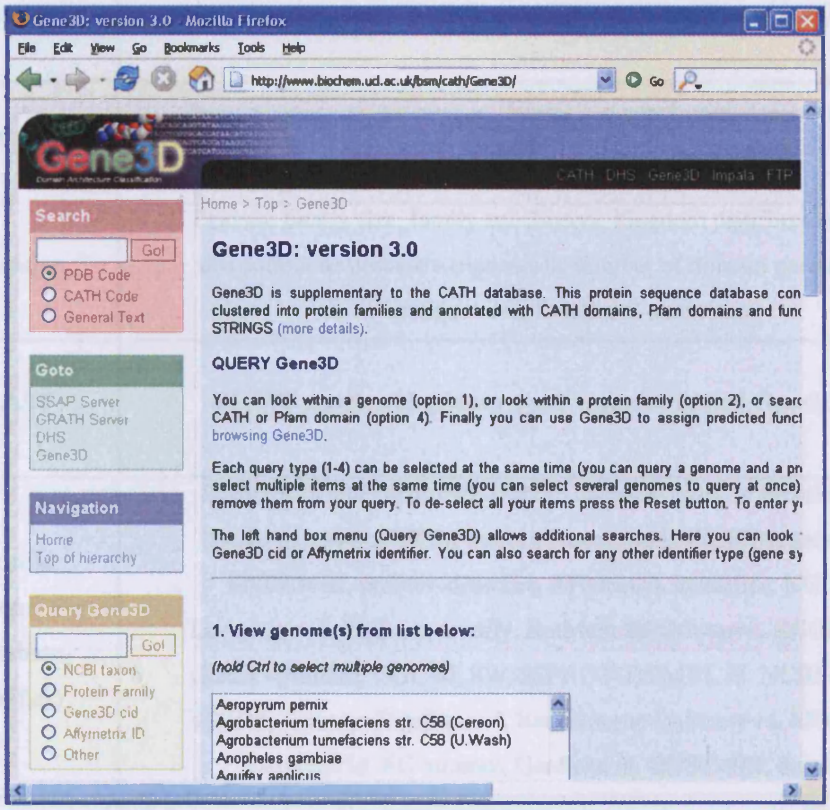


**Figure 2.16 Gene3D Website.** *Screenshot of user interface start page for Gene3D (www.biochem.ucl.ac.uk/bsm/cath/Gene3D_v3.0/gene3d.html). The interface for release 3 has since been further improved with the latest release 4 of Gene3D (bsmmac1.biochem.ucl.ac.uk:8080/Gene3D/).*

The user interface web pages run live database queries, the results of which are returned to the user as web page displays. The Gene3D website (shown in figure 2.16) has had over 11,000 web accesses, up to 868 visitors per month, the majority of which enter via the CATH or InterPro websites. The user interface consists of seven main queries, summarised in table 2.3.

**Table 2.3 User Interface Queries in Gene3D.** *For each user query the data returned is shown.*

| | Query Type | Data |
|---|---|---|
| 1 | Genome | Genome coverage, residue coverage |
| 2 | Protein Family | Number, kingdom distribution, domain architectures of family members |
| 3 | Protein | Protein sequence, description, family and subcluster, domain assignments, domain architecture, functional annotation |
| 4 | Domain | Domain family size, family subclusters, kingdom distribution, number of discontinuous domain assignments, number of domain partners, domain architectures, functional annotation |
| 5 | BLAST | BLAST search with protein query sequence to find Gene3D relatives |
| 6 | Specific Term (Database Identifier) | Search for Gene3D protein with 23 different types of protein identifiers (NCBI gi number, NCBI protein accession, ENSEMBL gene identifier, ENDEMBL protein identifier, Affymetrix identifier, KEGG entry, LocusLink id, FlyBase, Gadfly, RatMap, KEGG name, KEGG definition, KEGG position, COG id, SWISSPROT-TrEMBL id, NCBI description, OMIM, Genome DataBase id, Rat Genome Database id, Mouse Genome Database id, EC number, GenScan id, ENSEMBL description) |
| 7 | Functional Prediction | BLAST search with protein query sequence to find nearest relative with a domain architecture in Gene3D |

Where possible, all returned pages allow the user to browse related data. For example, genome queries return genome coverage and residue coverage statistics in Gene3D, and also provide links to lists of all protein families and proteins occurring in the genome. These in turn provide links to domain assignment and domain architecture assignments, which are displayed as diagrams (shown in figure 2.17). Domain architecture diagrams are clickable, allowing the user to link to individual CATH and Pfam domain pages by clicking on each domain box.

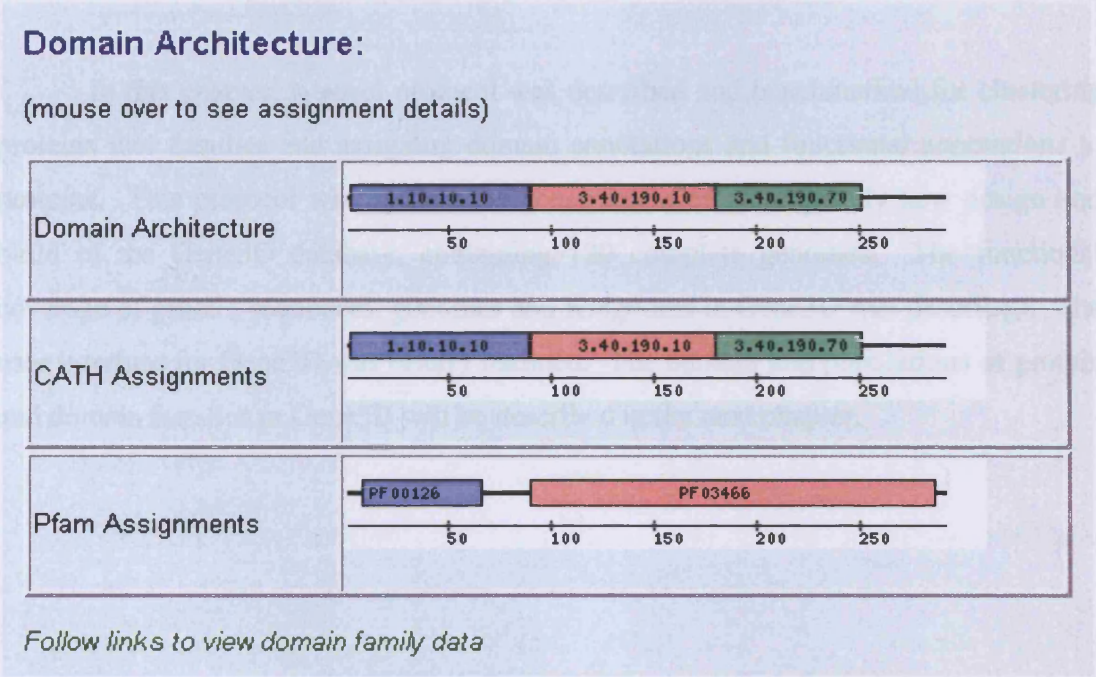


**Figure 2.17 Gene3D Domain Assignment Diagram.** *Screenshot of domain assignments (for Gene3 protein cid 00772590). Clickable domains (coloured boxes) link to source database websites.*

In addition to searching for specific identifiers from 23 different databases whose identifiers have been mapped to Gene3D proteins (for list of identifiers see table 2.3), the search facility and the functional prediction facility run a live BLAST comparison between the users input query protein sequence and all proteins in Gene3D to identify protein relatives in Gene3D. Users run ~150 BLAST searches per month.

## 2.6 Summary

In this chapter, a novel protocol was described and benchmarked for clustering proteins into families and assigning domain annotations and functional annotations to proteins. This protocol was used in the construction of a completely new design and build of the Gene3D database, containing 120 complete genomes. The functional coverage of protein sequences, genomes and Kingdoms in Gene3D was described. The user interface for Gene3D was briefly outlined. The number and populations of protein and domain families in Gene3D will be described in the next chapter.

# CHAPTER THREE

# Analysis of Protein Families and Domain Families in 120 Complete Genomes

## 3.1 Introduction

### 3.1.1 Power Laws in Protein Family Data

There are a number of resources that cluster large numbers of protein sequences into families. Recent analyses reveal that TRIBES clusters 83 complete genomes (311,257 protein sequences) into between 60,934 and 82,692 protein families, depending on the clustering granularity used. SYSTERS clusters 1,168,542 proteins from SWISSPROT, TrEMBL and 11 complete genomes into 158,153 protein families of which 110,322 are singleton families containing only a single protein sequence.

Comparison between different protein family resources that use different methodologies to cluster proteins can be difficult. However, some general characteristics of large scale protein family clustering can be observed, notably the observation of power law distributions in family cluster data. A power law describes the domination of a population by a selected few. Power law relationships were described in economic theory in the 19$^{th}$ century by Vilfredo Pareto illustrating the relationship whereby 20% of the total population earns 80% of the total income. A Pareto or power law distribution is a probability distribution where the density is proportional to a power function $P(x) = 1/Zx^{alpha}$ for any real value alpha and normalisation factor Z. In biological data there are many established examples of power law distributions (Luscombe *et al.*, 2002). In terms of protein families, the frequency of protein families of a particular size (protein families containing a certain number of members) has been reported to follow a power law relationship where most families are very small containing few protein relatives, while a few families are very large having many protein relatives, in both clustering of completely sequenced genomes and in clustering SWISSPROT-TrEMBL protein sequences (Enright *et al.* 2003; Kunin *et al.*, 2005; Meinel *et al.*, 2003). Whilst Luscombe *et al.*, report that two other functions (triple-exponential and lognormal) also describe these distributions quite well, they

conclude that the power law distribution is a better descriptor of genomic data, since power law functions fit many different biological distributions with a more simple function compared to triple-exponential and lognormal functions.

In a genomic context, the protein families within an individual genome also adopt power law behaviour whereby a small number of protein families are large and have many relatives throughout the genome, but the vast majority of protein families are small having few relatives within a genome (Harrison & Gerstein, 2002).

### 3.1.2 Novel Protein Families

Kunin *et al.*, (Kunin *et al.*, 2003) have investigated the rate of discovery of novel protein families as successive completed genome sequences are released. Interestingly they calculate that the number of novel protein families identified over time has remained constant, indicating that our coverage of protein family sequence space is not yet saturating.

Whilst the phylogenetic position of a newly completed genome determines the number of novel protein families identified (for example a closely related strain of a previously sequenced genome contains less novel protein families than a newly sequenced genome with no previously sequenced closely related genome), the recent addition of several eukaryotic genomes does not account for the observed trend. Excluding eukaryotes, Kunin *et al.*, find that protein family sequence space occupied by prokaryotic genomes is still being expanded at a constant rate, that is novel protein families are just as likely to be identified in the average newly sequenced prokaryotic genome today. Indeed, these novel families are not only genome specific but are likely to remain very small until closely related genomes are sequenced to identify close relatives. This contrasts with large protein families with relatives in many different genomes, where each newly sequenced genome adds more relatives to the protein family.

### 3.1.3 Domain Assignment to Genomes

Genome coverage in genomic domain assignment resources can be defined as the average percentage of genes in a genome that have at least one domain assignment. There are two resources that assign domains to genomic sequences where genome coverage information is available, the SUPERFAMILY database and the Genomic Threading Database; both assign SCOP domains to complete genome sequences.

Amongst the 220 genomes in SUPERFAMILY (described previously, see section 2.1.6), as of February 2005, the genome coverage (residue coverage in parenthesis) ranges from 19% (15%) to 81% (71%), with an average coverage of 57.4% (49.5%).

In the Genomic Threading Database (described previously, see section 2.1.6), as of February 2005, the genome coverage (residue coverage in parenthesis) ranges from 46.9% (37.7%) to 97.2% (79.1%), with an average coverage of 81.6% (61.6%) in the 218 genomes within the database.

Both these databases are able to achieve impressive genome coverage with domain assignments in over half the proteins within an average genome.

#### 3.1.3.1 Un-assignable Regions

When considering how much of a genome is described by domain assignments to proteins, and the limits of this characterisation, domain assignment coverage by percentage of total protein residues is found to be lower than coverage by percentage of total proteins with at least one domain assignment, since a significant proportion of residues cannot be assigned a domain. These un-assignable residues include signal peptides (SP), transmembrane helix (TM), disordered regions with no regular secondary structure (NORS), coiled-coil (CC) or low-complexity (LC) regions.

Liu and Rost (Liu and Rost, 2002) described the percentage of genes in 30 genomes (6 archaea, 20 bacteria, 4 eukaryota) that have been identified as belonging to some of these groups, and therefore cannot be assigned a domain. Liu and Rost

reported the percentage of total proteins containing transmembrane regions (22% of proteins had at least one TM region, and that half of these contained more than five TM regions), coiled-coils (8%) and NORS regions (16%). The distribution across individual genomes differed by Kingdom: twice as many eukaryotic proteins than prokaryotic contained coiled-coil regions and almost eight times as many eukaryotic proteins contained NORS regions than prokaryotic proteins, although the percentage of proteins with a transmembrane region was similar across all Kingdoms. The total percentage of proteins that contained regions that cannot be assigned a domain ranged from 30-40%. Recent analyses of 203 completely sequenced genomes has shown that, on average, the percentage of residues in a genome that are un-assignable is 7.5%, contrasting with the percentage of genes in a genome containing these un-assignable regions of 16.6% (Russell Marsden, personal communication).

### 3.1.4 Domain Architecture

Vogel *et al.* (Vogel *et al.*, 2004) introduced the term 'domain architecture' to describe the complete domain makeup of a protein as the string of known SCOP domains and un-assigned regions assigned to a protein. In 261,344 multi-domain proteins from 131 genomes they identify 28,387 different domain architectures. They show that proteins sharing identical domain architectures tend to have similar functions and that this relationship is domain order dependent since this is not the case if domain order is swapped. Domain architectures found in different proteins are likely to come about by duplication from a common ancestor. Thus proteins with the same domain architecture can be regarded as belonging to a single protein family and sharing a common evolutionary ancestor.

## 3.2 Objectives

In this chapter the distribution of protein and domain families in different Kingdoms is explored and the consistency and characterisation of domain architectures in Gene3D is examined. Domain family assignments and protein family information is used to investigate the Kingdom distribution of protein families and domain families in 120 genomes. Finally, the sequence diversity of protein families is described, and functional annotation in Gene3D is used to characterise sequence diverse and sequence invariant protein families.

## 3.3 Results

### 3.3.1 Analysis of Protein Family Populations in Gene3D

A total of 854,897 proteins from 120 completely sequenced genomes in Gene3D were clustered into protein families and subclusters according to the PFscape protocol. The number of protein families and subclusters can be seen in table 3.0 below.

**Table 3.0 Number of Protein Families and Subclusters in Gene3D.** *Clustering levels of protein family and subclusters of 35%, 60%, 95% and 100% sequence identity are shown for all clusters and for non-singleton clusters.*

| Cluster Level | Number of Clusters | |
|---|---|---|
| | Total Clusters | Non-Singleton Clusters |
| Protein Family | 112,464 | 50,219 |
| S35 subcluster | 228,253 | 166,008 |
| S60 subcluster | 356,392 | 294,147 |
| S95 subcluster | 459,675 | 397,430 |
| s100 subcluster | 501,135 | 438,890 |

Using a threshold BLAST E-value for matches below 0.0001 (the default threshold used in granularity benchmarking and TRIBES, described previously, see section 2.4.1) resulted in 417,160,739 significant similarities between all the proteins. A small fraction of proteins did not produce any significant BLAST E-values when compared to all other proteins. These protein sequences, mostly less than 25 residues in length, are highly unlikely to produce a significant BLAST E-value, even when compared to themselves since these sequence are not long enough to produce significant BLAST alignments. Such sequences were left as singleton clusters consisting only of themselves. This does not necessarily indicate that the protein has no relatives, but merely that a relative could not be identified by BLAST similarity.

#### 3.3.1.1 Size Distribution of Protein Families

After the clustering process was completed 112,464 protein families had been identified. 62,245 of these gene families are singleton gene families containing only a

single member, leaving 50,219 non-singleton gene families. The size of these non-singleton protein families exhibits power law like behaviour whereby 20% of protein families contain 70% of protein sequences. This distribution appears as a linear relationship when plotted on double-logarithmic axes. This observation is in agreement with several previous analyses (Enright *et al.*, 2002; Luscombe *et al.*, 2003; Kunin *et al.*, 2005). A small percentage (1%) of protein families (823 families) are very large, containing more than 100 relatives and accounting for a large percentage (25%) of protein sequences. Conversely, a large percentage of protein families (79%) are very small, containing less than ten relatives and accounting for only 27% of total proteins. As figure 3.0 shows no obvious deviation from a power law distribution, it can be concluded that the protein family clustering process employed by Gene3D does not artificially over-represent protein families of any particular size.

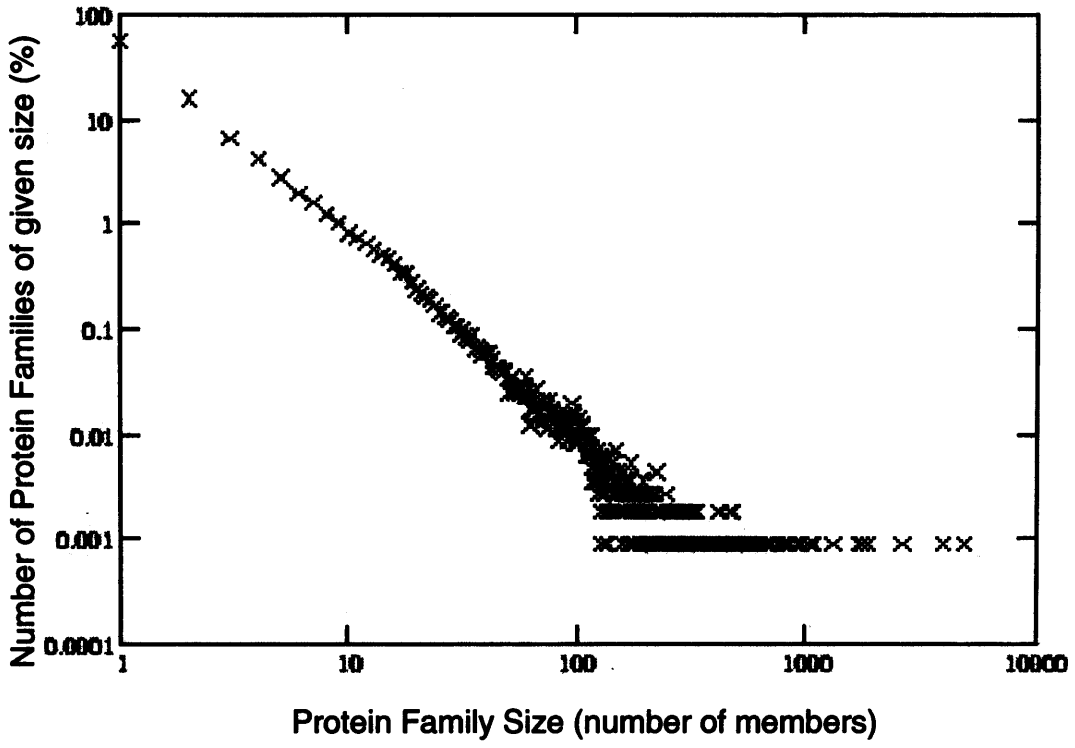


**Figure 3.0 Power Law Distribution of Protein Families.** *Number of protein families of a given size (plotted as a percentage) against protein family size on double log axes approaches power-law behaviour. A power-law is of the form* $y = ax^{-b}$ *and appears as a straight line when plotted on double log axes.*

The largest protein families are shown in table 3.1 below. The largest protein family comprises zinc finger containing transcription regulators with 4,842 members. These large families have been identified by previous analyses (SYSTERS, Meinel *et al.*, 2005; TRIBES, Enright *et al.*, 2003) and contain proteins that are performing important generic functions, such as regulation of transcription, signal transduction and DNA replication exploited by organisms in all kingdoms. When divided into prokaryotic and eukaryotic protein families, the largest protein families in prokaryotes are involved in metabolism and transcription regulation, whereas in eukaryotes the largest families are involved in regulation of transcription, G-protein coupled receptor signal transduction pathways and cell adhesion.

**Table 3.1 Largest Protein Families in Gene3D.** *Top ten largest protein families with the number of members shown for each family. Protein family names are derived from the most common GO term assigned to family members.*

| Protein Family Name | Size (# Relatives) |
|---|---|
| Nucleic Acid Binding (Zinc Finger) | 4,842 |
| ATP Binding (ABC Transporter) | 3,969 |
| Rhodopsin-like Receptor Activity | 2,638 |
| Oxidoreductase Activity (NAD(P)-Binding) | 1,741 |
| Protein Serine/Threonine Kinase Activity | 1,732 |
| Trypsin Activity | 1,309 |
| DNA Binding | 1,074 |
| Protein-tyrosine Kinase Activity | 1,030 |
| Kinase Activity | 1,029 |
| ATP Binding | 968 |

### 3.3.1.2 Diversity of Protein Families in Gene3D

The number of s35 subclusters within a protein family can be used as a measure of the sequence diversity within the protein family. Family diversity can be measured by dividing number of s35 subclusters by the protein family size (the number of relatives in the family). The distribution of protein family diversity can be plotted as shown in figure 3.1. This distribution shows that the average diversity for a protein family is 0.225. Protein families with a diversity less than 0.05 can be defined as sequence invariant, whilst those with a diversity greater than 0.5 can be defined as sequence diverse. Functional classification of diverse and invariant protein families was

determined by protein family GO, KEGG, COG functional assignments and Pfam domain assignments. These are discussed below.

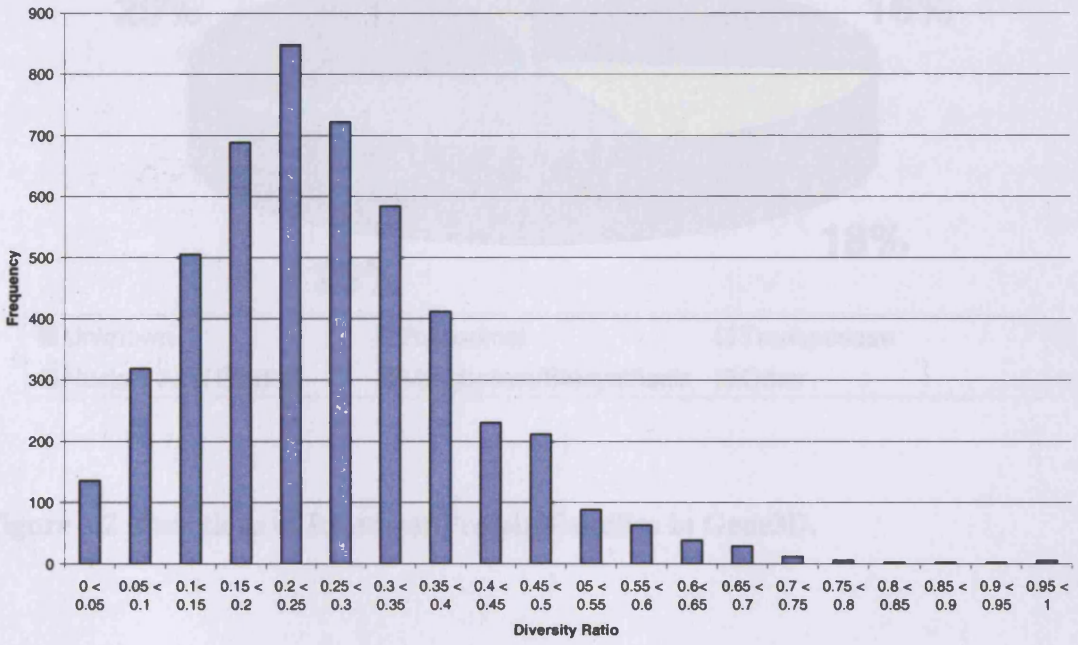


**Figure 3.1 Frequency Distribution of Protein Family Diversity.** *Number of protein families with a given diversity for protein families with at least 20 members.*

3.3.1.2.1 Function of Invariant Protein Families

The function of invariant protein families is shown in figure 3.2. The majority of sequence invariant protein families comprise proteins involved in two specific core processes: nucleic acid binding proteins involved in nucleotide excision/repair processes; and metabolism and biosynthesis enzymes involved in the tricarboxylic acid cycle, purine/pyrimidine biosynthesis and amino acid biosynthesis. Many of the largest invariant protein families are transposases, proteins necessary for efficient DNA transposition, many of which bind metal ions required for catalysis of DNA cleavage at specific sites. The next largest group of sequence invariant protein families are ribosomal proteins, large ribonucleoprotein particles required for translation of mRNA into protein in both prokaryotes and eukaryotes.

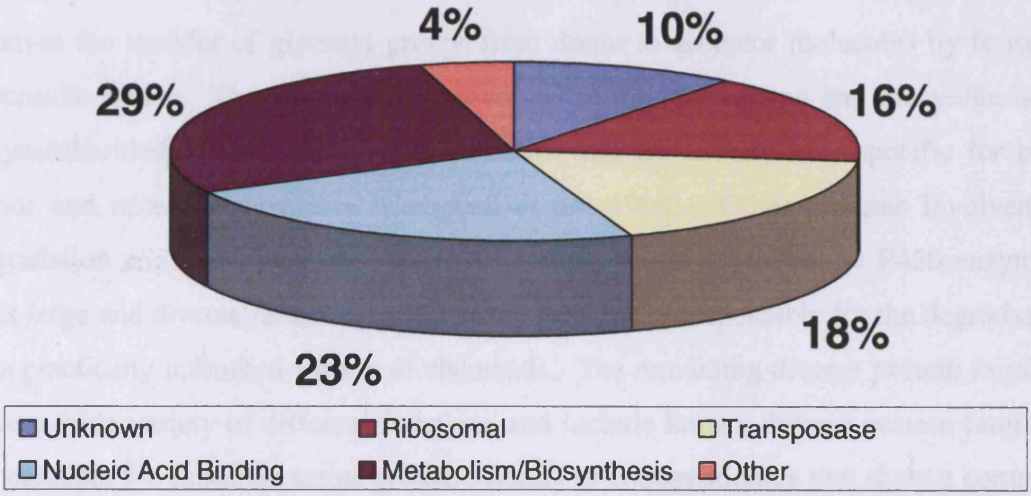


**Figure 3.2 Functions of Invariant Protein Families in Gene3D.**

3.3.1.2.2 Function of Diverse Protein Families

The vast majority of diverse protein families are poorly functionally annotated, as shown in figure 3.3. Those protein families that have been assigned a function perform many different functional roles. Functional classification of these protein families reveals only a few large functional groups. The largest group of diverse protein families is the two component sensor histidine kinases. These bacterial proteins combine signal recognition, signal transduction and gene activation in a two protein system. The sensor histidine kinase interacts directly with a signal ligand, or a receptor that binds the signal ligand. Variability in signal ligand, or with specific response regulator proteins that bind DNA, thus activating transcription, provide a wide range of virulence factors and antimicrobial resistance responses in pathogenic bacteria and fungi, as well as regulation of essential cellular functions. Diversity in sensor histidine kinases is required to detect an immense diversity of possible signal ligands.

The next largest groups of diverse protein families comprise methyl-accepting chemotaxis proteins and glycosyltransferases. Bacterial chemotactic signal transducer proteins respond to changes in the environmental concentration of a wide range of attractants and repellents and transduce a signal from the outside to the inside of the cell

in response, via deamidation and reversible methylation. Attractants increase the level of methylation, whilst repellents decrease it. Glycosyltransferases are proteins that catalyse the transfer of glycosyl groups from donor to acceptor molecules by forming glycosidic bonds. These proteins are involved in the degradation and biosynthesis of polysaccharides, glycoproteins and glycolipids and are usually very specific for both donor and acceptor substrates (Campbell *et al.*, 1998). Other proteins involved in degradation are found in diverse protein families, notably cytochrome P450 enzymes. This large and diverse family of enzymes are principally responsible for the degradation of a practically unlimited variety of chemicals. The remaining diverse protein families have a wide variety of different functions and include known diverse protein families, for example the subtilisin serine protease family of endopeptidases that share a common catalytic triad but possess highly diverse N and C terminal extensions.

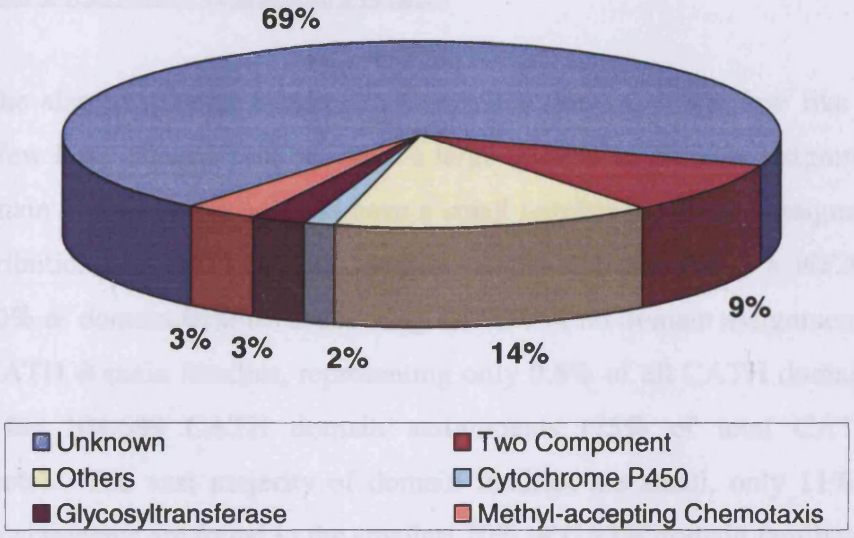


**Figure 3.3 Functions of Diverse Protein Families in Gene3D.**

Both highly diverse and highly invariant protein families perform functions critical to an organism's survival. The diversity of these protein families is dictated by their biological function. Highly invariant protein families perform very specific functions and interact with only a few ligands associated with core metabolic and nuclear processes. Highly diverse protein families perform a myriad of functions,

interacting with a huge variety of ligands regulating complex cellular responses, often to environmental stresses.

### 3.3.2 Analysis of Domain Family Populations in Gene3D

Scanning all protein sequences in Gene3D against a library of HMMs representing CATH structural domains and Pfam sequence domains was undertaken according to the PFscape protocol. Using a threshold HMM E-value for matches of 0.01, an HMM percentage model matched cut-off of 50% and a domain overlap cut-off of 30% resulted in 417,132 significant CATH domain assignments and 508,348 significant Pfam domain assignments.

#### 3.3.2.1 Size Distribution of Domain Families

The size of domain families in Gene3D follows a power law like behaviour, where a few large domain families have a large number of domain assignments whilst most domain families are small and have a small number of domain assignments. The size distribution of CATH domain families closely follows Pareto's 80/20 Law, the largest 20% of domain families accounting for 80% of all domain assignments. The ten largest CATH domain families, representing only 0.8% of all CATH domain families) account for 104,689 CATH domain assignments (25% of total CATH domain assignments). The vast majority of domain families are small, only 11% of CATH domain assignments are found in the smallest 70% of CATH domain families.

The sizes of CATH and Pfam domain families are shown in figure 3.4 below. There are 91,216 more Pfam domain assignments than CATH domain assignments in total. These additional Pfam domain assignments are mainly from medium sized Pfam domain families containing between 10 and 200 domain family members, which are not related to domain families represented in CATH.

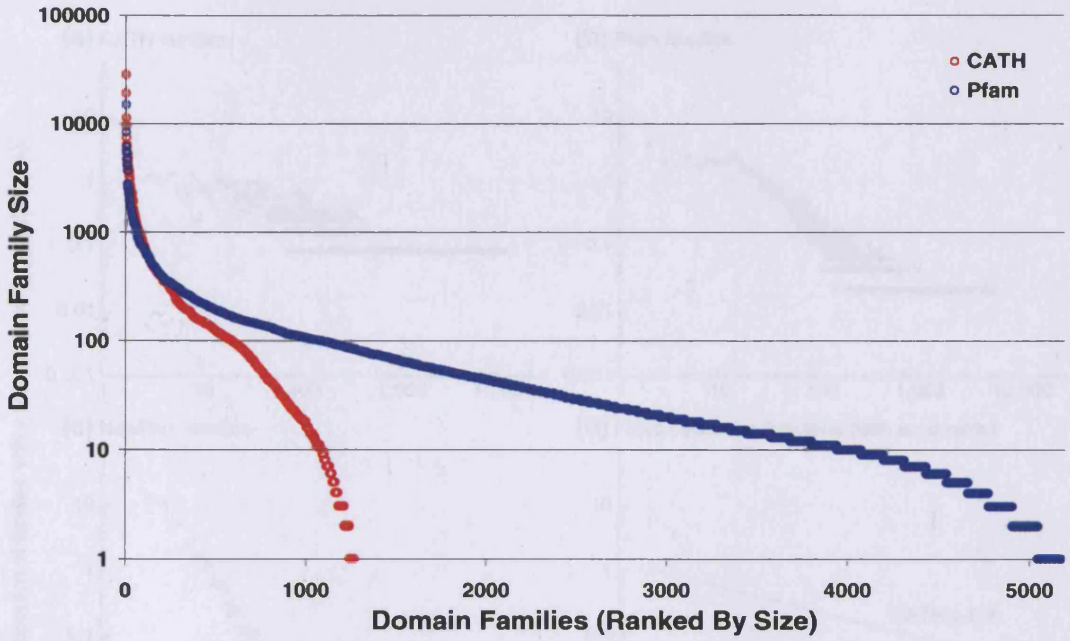


**Figure 3.4 Sizes of Domain Families.** *Domain family size (log scale) against domain families ranked by size (number of domains assigned to family) for CATH homologous superfamilies (red) and Pfam domain families (blue).*

The s35 sequence family size distribution of CATH, Pfam and uncharacterised (Newfam) domain families also shows power law like behaviour. This is illustrated in the power law plots in figure 3.5 below for CATH, Pfam and Newfam families and show that most of the uncharacterised families (Newfam) tend to be much smaller than the CATH and Pfam families.

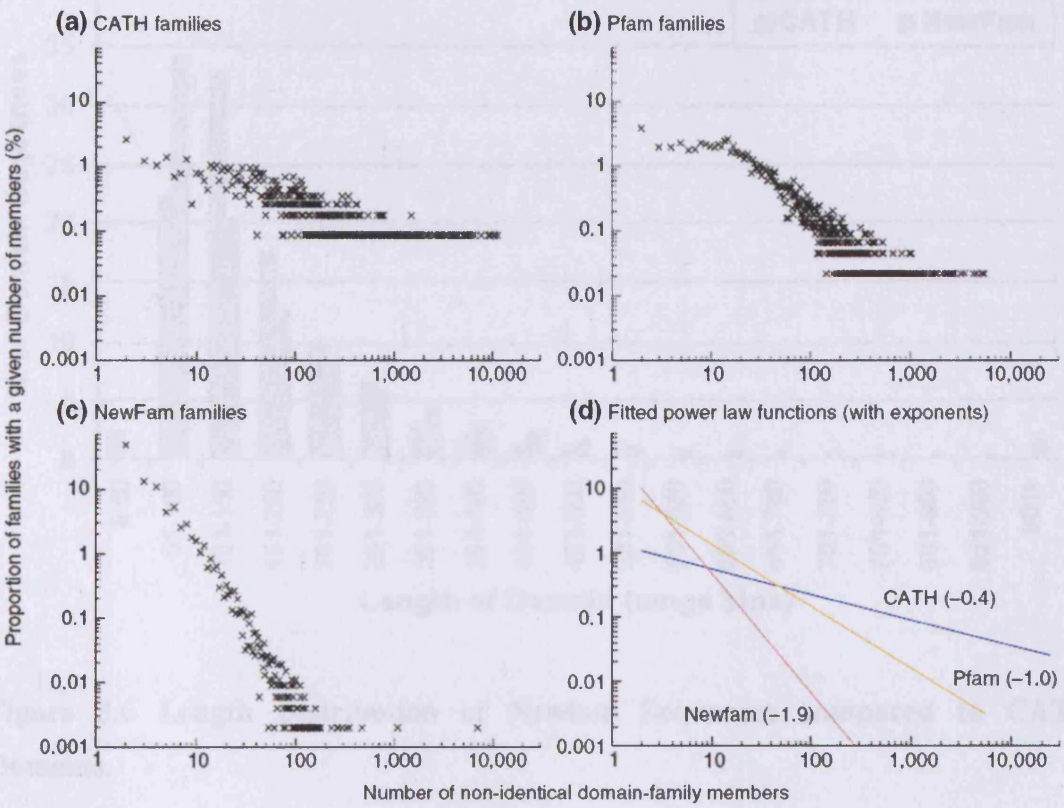


**Figure 3.5 Log-log Plots of (a) CATH, (b) Pfam and (c) Newfam Families.** *Graphs show power-law like behaviour. Fitted power law functions and their exponents are shown for comparison (d). Note that most Newfam families are small with relatively few members.*

Although the Newfam families are much smaller families, with less members than CATH and Pfam families, it is encouraging to note that the length distribution of Newfam sequences is close to the length distribution of domains classified in the CATH (see figure 3.6 below). This is indicative that many Newfam families, whilst sparsely populated, are likely to represent protein domains. A small percentage of Newfam domains are very long and are thus more likely to represent multiple domains.

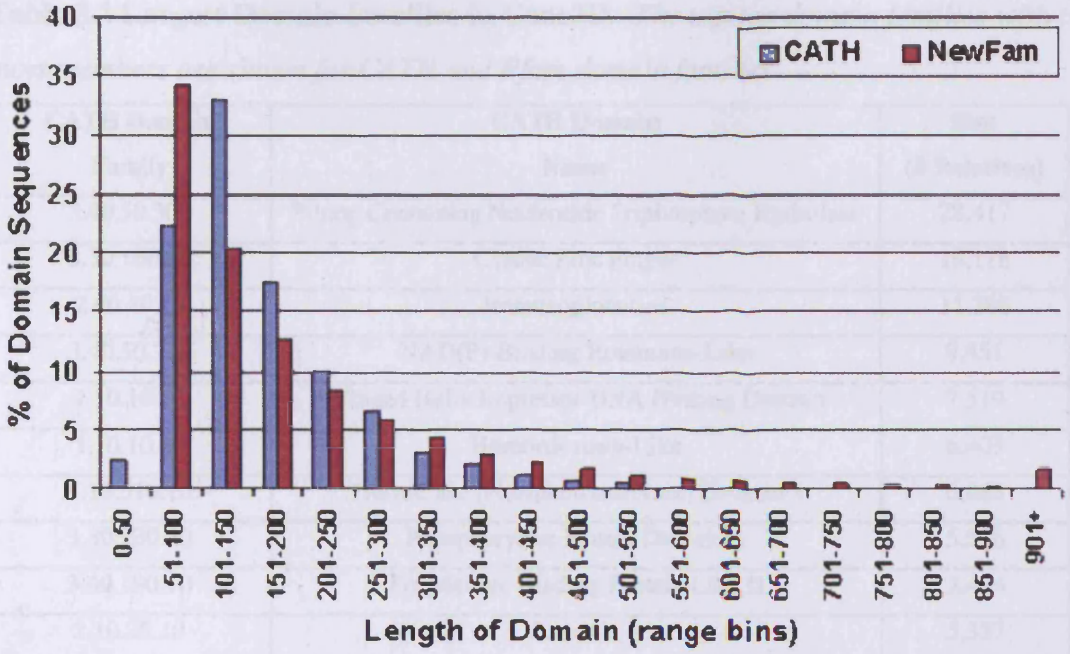


**Figure 3.6 Length Distribution of Newfam Sequences compared to CATH Domains.**

The number of domain assignments for the largest CATH and Pfam domain families are shown in table 3.2 below. These largest families identified in Gene3D have also been identified by previous analyses (Buchan *et al.*, 2002; Bateman *et al.*, 2004).

**Table 3.2 Largest Domain Families in Gene3D.** *The top ten domain families with the most members are shown for CATH and Pfam domain families.*

| CATH Domain Family | CATH Domain Name | Size (# Relatives) |
|---|---|---|
| 3.40.50.300 | P-loop Containing Nucleotide Triphosphate Hydrolase | 28,417 |
| 3.30.160.60 | Classic Zinc Finger | 19,118 |
| 2.60.40.10 | Immunoglobulins | 11,386 |
| 3.40.50.720 | NAD(P)-Binding Rossmann-Like | 9,451 |
| 1.10.10.10 | Winged Helix Repressor DNA Binding Domain | 7,519 |
| 1.10.10.60 | Homoedomain-Like | 6,403 |
| 1.10.510.10 | Transferase (Phosphotransferase) Domain 1 | 6,088 |
| 3.30.200.20 | Phosphorylase Kinase Domain 1 | 5,526 |
| 3.40.190.10 | Periplasmic Binding Protein-Like II | 5,454 |
| 2.10.25.10 | Laminin | 5,327 |

| Pfam Domain Family | Pfam Domain Name | Size (# Relatives) |
|---|---|---|
| PF00096 | Zinc Finger, C2H2 Type | 15,012 |
| PF00005 | ABC Transporter | 8,352 |
| PF00069 | Protein Kinase Domain | 6,085 |
| PF00028 | Cadherin Domain | 5,601 |
| PF00041 | Fibronectin Type III Domain | 5,389 |
| PF00400 | WD Domain, G-Beta Repeat | 5,278 |
| PF00001 | Seven Transmembrane Receptor (Rhodopsin Family) | 4,709 |
| PF00023 | Ankyrin Repeat | 4,679 |
| PF00083 | Sugar (and Other) Transporter | 4,252 |
| PF00076 | RNA Recognition Motif (RRM, RBD, RNP Domain) | 4,079 |

### 3.3.3 Domain Assignments to Protein Families in Gene3D

By combining protein family and domain assignment data in Gene3D, the domain coverage of protein families can be characterised. The figures below illustrate the extent of CATH (figure 3.7) and Pfam (figure 3.8) domain assignments to protein families in Gene3D.

The graphs show that for most protein families with more than 5 members, at least half of the protein families have a Pfam domain assignment. The coverage of

protein families is slightly less comprehensive for CATH domains, where for protein families with at least 20 members, at least half of the protein families have a CATH domain assignment. Coverage of protein families by CATH is lower than that of Pfam. Additionally, families annotated with CATH are, on average, much larger than families annotated with Pfam, this is because small families are under-represented in CATH, whilst Pfam annotates many more smaller families. Structural data allows detection of very distant relatives, so that Pfam families are often merged once representative structures are solved. This can be clearly seen in Figure 3.7 and 3.8 below.

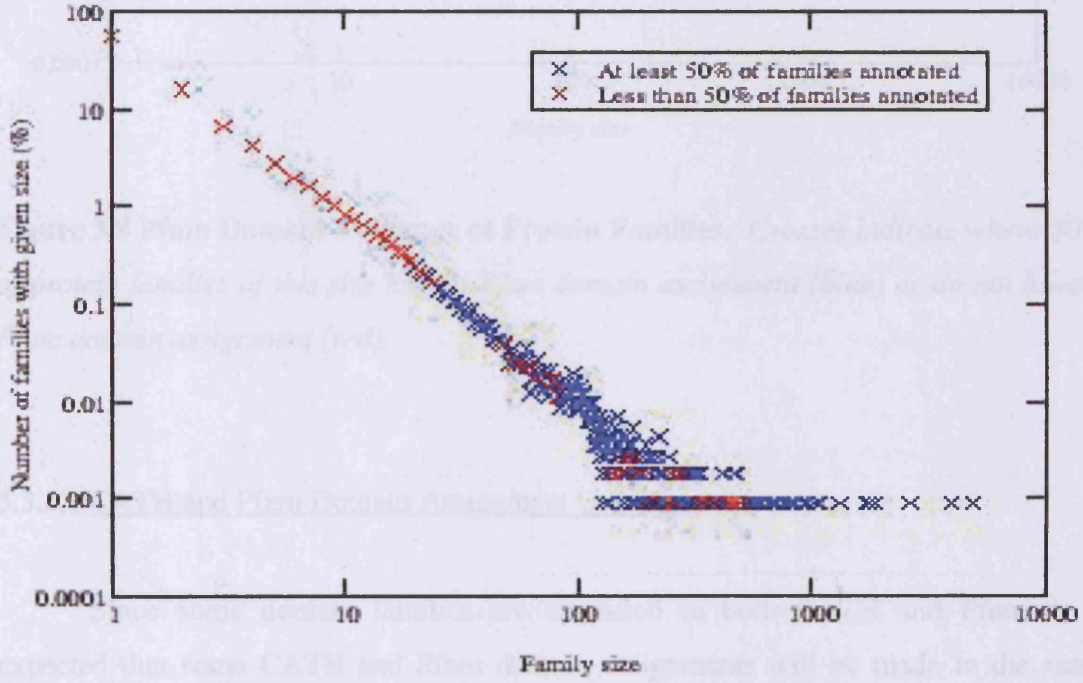


**Figure 3.7 CATH Domain Coverage of Protein Families.** *Crosses indicate where 50% of protein families of this size have a CATH domain assignment (blue) or do not have a CATH domain assignment (red).*

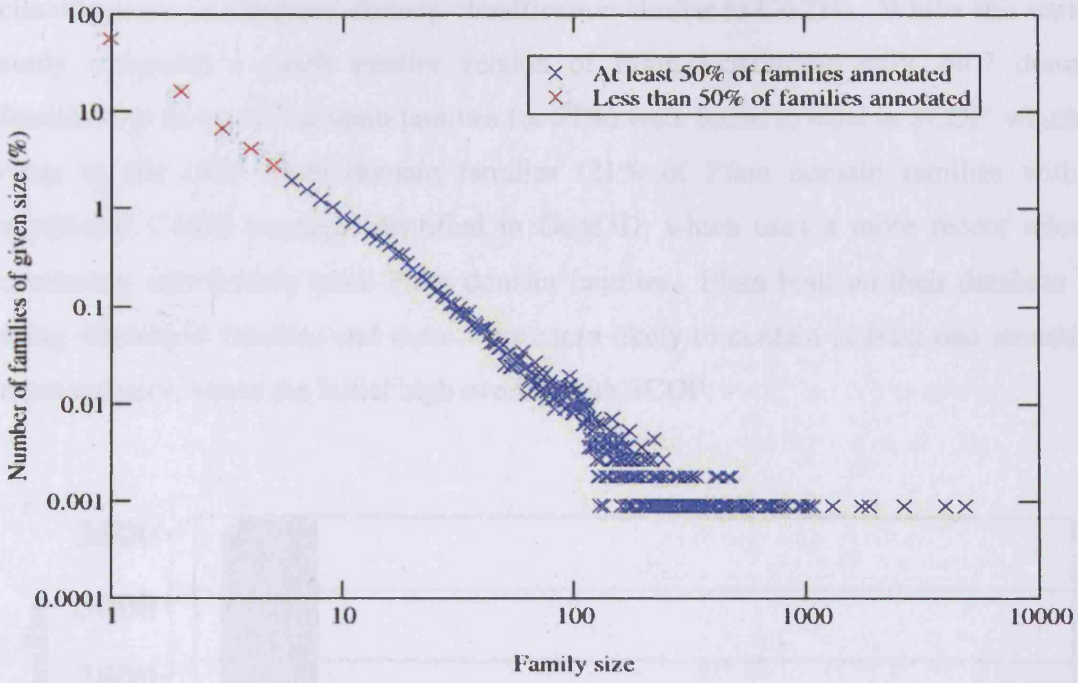


**Figure 3.8 Pfam Domain Coverage of Protein Families.** *Crosses indicate where 50% of protein families of this size have a Pfam domain assignment (blue) or do not have a Pfam domain assignment (red).*

3.3.3.1 CATH and Pfam Domain Assignment Overlap

Since some domain families are classified in both CATH and Pfam, it is expected that some CATH and Pfam domain assignments will be made to the same regions within protein sequences and result in overlapping CATH and Pfam domain assignments. Figure 3.9, below, shows the extent of the overlap between CATH and Pfam domain assignments. As figure 3.9 shows, 21% of Pfam domain families have a significant overlap (where, on average, greater than 80% of the Pfam domain assignment is overlapped by a CATH domain assignment) with CATH domain families, suggesting that these Pfam and CATH families may be regarded as equivalent domain families. Whilst 68% of the 5,179 Pfam domain families assigned in Gene3D have no significant overlap (i.e. overlap less than or equal to 20%) with any CATH domain families, 11% of Pfam domain families have an intermediate level of overlap, suggesting that these Pfam domain families may comprise additional structurally uncharacterised domains. This is in agreement with an earlier study by Elofsson and Sonnhammer (Elofsson and Sonnhammer, 1999), which compared Pfam and SCOP

classifications (a structural domain classification similar to CATH). Whilst this earlier study compared a much smaller version of Pfam (containing only 1407 domain families), of these, 802 domain families (or 57%) were found to exist in SCOP, which is close to the 1087 Pfam domain families (21% of Pfam domain families with a significant CATH overlap) identified in Gene3D, which uses a more recent release containing significantly more Pfam domain families. Pfam built up their database by using the largest families and these were more likely to contain at least one structural representative, hence the initial high overlap with SCOP.

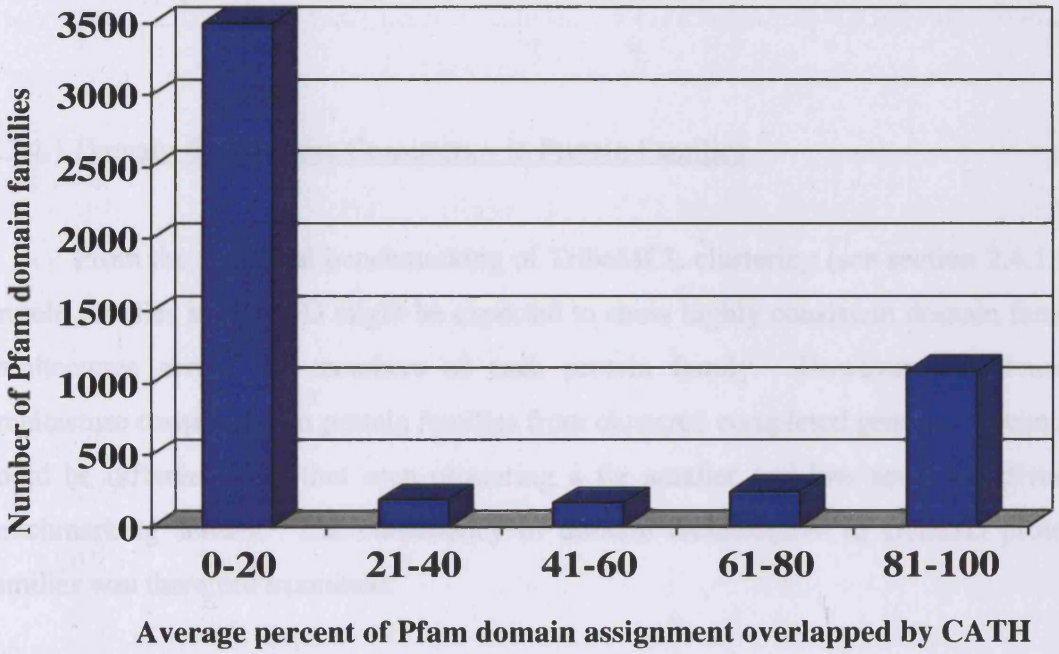


**Figure 3.9 Overlap of CATH and Pfam Domain Assignments.** *Most Pfam families show negligible overlap with CATH (0-20% sequence overlap) but 1,085 Pfam families show substantial overlap (81-100% sequence overlap). There are relatively few Pfam families with an intermediate level of overlap.*

### 3.3.4 Domain Architectures in Gene3D

CATH and Pfam domain assignments were combined into domain architectures using the DomainFinderII protocol described previously (see section 2.4.2.8). When combining CATH and Pfam domain assignments into domain architectures, 63% of all Pfam domain assignments are excluded by an overlapping CATH assignment. As

figure 3.9 shows, there are 1,085 Pfam domain families that overlap significantly (greater than 80%) with CATH. Since these 1,085 families represent 63% of all Pfam assignments, they are large domain families, where the domain family is also represented in the CATH classification.

Domain architectures were assigned to a total of 386,340 proteins in Gene3D. 59% of these domain architectures contain only CATH domain assignments, 32% contain only Pfam domain assignments, leaving 9% of domain architectures containing both CATH and Pfam domain assignments. This is consistent with the level of observed overlap between CATH and Pfam domain classifications.

#### 3.3.4.1 Domain Architecture Consistency in Protein Families

From the structural benchmarking of TribeMCL clustering (see section 2.4.1.1), protein families in Gene3D might be expected to show highly consistent domain family architectures across the members of each protein family. However, the domain architecture consistency in protein families from clustered completed genome sequences could be different from that seen clustering a far smaller and less sequence diverse benchmarking dataset. The consistency of domain architectures in Gene3D protein families was therefore examined.

Figure 3.10 shows domain architecture consistency in 7,453 protein families which contain at least 3 relatives, in which the domain architecture covers on average at least 80% of the protein's residues. 77% (5,719) of these protein families have >90% of annotated members with the same domain architecture, whilst 87% of the protein families have more than 70% of members sharing a common domain architecture. Although these 7,453 protein families represent only ~11% of non-singleton protein families in Gene3D, comprising 17% of non-singleton sequences, they give some indication of consistency in domain architecture achieved by the TribeMCL clustering.

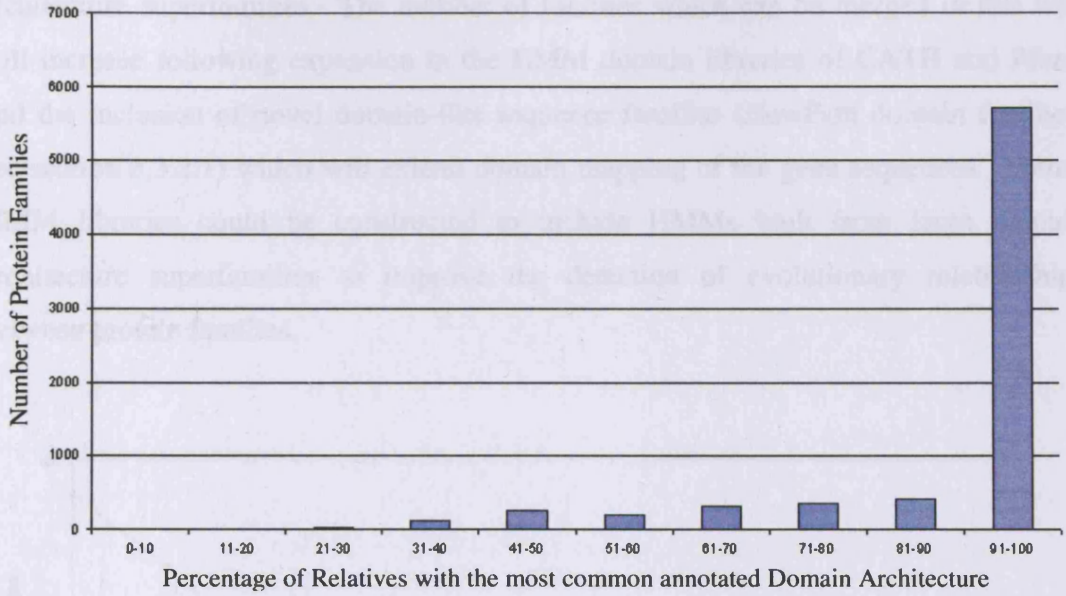


**Figure 3.10 Consistency of Domain Architecture in Protein Families.** *More than three-quarters of protein families have more than 90% of their members annotated with the same domain architecture.*

3.3.4.2 Domain Architecture Superfamilies in Gene3D

For protein families that are completely annotated, domain architecture information can be used to resolve subgroups of dissimilar relatives into separate families sharing common architectures. It also allows mapping between protein families with common domain architectures. Figure 3.11 shows the distribution of 2,212 domain architectures that are found in 5,719 protein families, which contain at least 3 relatives and in which the domain architecture covers on average at least 80% of protein residues, and where >90% of annotated protein family members have the same domain architecture. 60% of these domain architectures are unique to a single protein family, a further 20% only occur in two protein families, whilst the remaining 20% occur in 3 or more protein families.

This suggests that TribeMCL is rather conservative at clustering sequences, but this is desirable since it preserves the consistency of the domain architecture within a protein family. Sequences sharing common domain architectures but placed in separate protein families can be merged into the same domain architecture superfamily. For example, the 5,719 protein families described above can be collapsed into 2,212 domain

architecture superfamilies. The number of families which can be merged in this way will increase following expansion in the HMM domain libraries of CATH and Pfam; and the inclusion of novel domain-like sequence families (NewFam domain families, see section 3.3.2.1) which will extend domain mapping of the gene sequences. Future HMM libraries could be constructed to include HMMs built from large domain architecture superfamilies to improve the detection of evolutionary relationships between protein families.

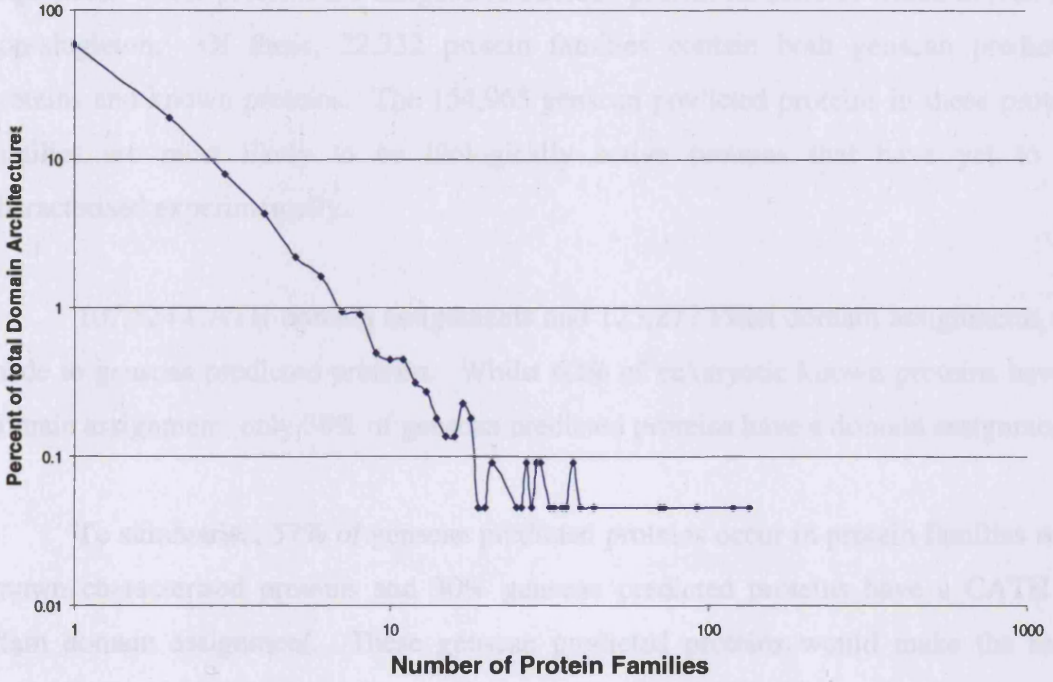


**Figure 3.11 Domain Architecture Distribution across Protein Families.** *The percentage of domain architectures against the number of different protein families in which they are identified. Note that 60% of complete domain architectures are found in only a single protein family.*

## 3.3.5 Using Gene3D Families to Validate Genscan Predictions

Genscan protein sequences are in-silico translations of open reading frames identified in eukaryotic genomic DNA sequences using the Genscan (Burge and Karlin, 1997) prediction program. These predicted protein sequences are not included in analysis in Gene3D since their validation as biologically active entities cannot be

verified from existing sources, but have been included in the PFscape process to expand the information for Tribe-MCL clustering. The distribution of these genscan predicted proteins within the protein families and domain families identified in Gene3D is of interest because it can indicate the biological relevance of these protein predictions and allow an estimation of the proportion of predicted proteins that are likely to have a biological function.

Of the 854,897 proteins in Gene3D, 270,846 proteins are genscan predicted sequences. These proteins are assigned to 127,804 protein families of which 29,765 are non-singleton. Of these, 22,332 protein families contain both genscan predicted proteins and known proteins. The 154,965 genscan predicted proteins in these protein families are most likely to be biologically active proteins that have yet to be characterised experimentally.

107,524 CATH domain assignments and 125,277 Pfam domain assignments are made to genscan predicted proteins. Whilst 62% of eukaryotic known proteins have a domain assignment, only 30% of genscan predicted proteins have a domain assignment.

To summarise, 57% of genscan predicted proteins occur in protein families with known characterised proteins and 30% genscan predicted proteins have a CATH or Pfam domain assignment. These genscan predicted proteins would make the most promising targets for expression studies in eukaryotes to determine whether they are expressed, as suggested by Gene3D, and should therefore be studied further.

### 3.3.6 Genome Coverage in Gene3D

Genome coverage can be assessed in two ways: Gene Coverage - the percentage of protein sequences in a genome that have at least one domain assignment; and Residue Coverage - the percentage of protein residues in a genome that are covered by one or more domain assignments. The first of these measures simply describes the percentage of proteins in a genome for which some domain information can be assigned, and does not necessarily provide an accurate measure of how completely such information describes these proteins as a whole. The percentage of residues within a

genome with domain assignments gives a much clearer indication of the completeness of protein annotation by domain assignment.

The following figures show gene coverage (figure 3.12) and residue coverage (figure 3.13) in all the 120 genomes in Gene3D, data for individual genomes can be seen in appendix I.

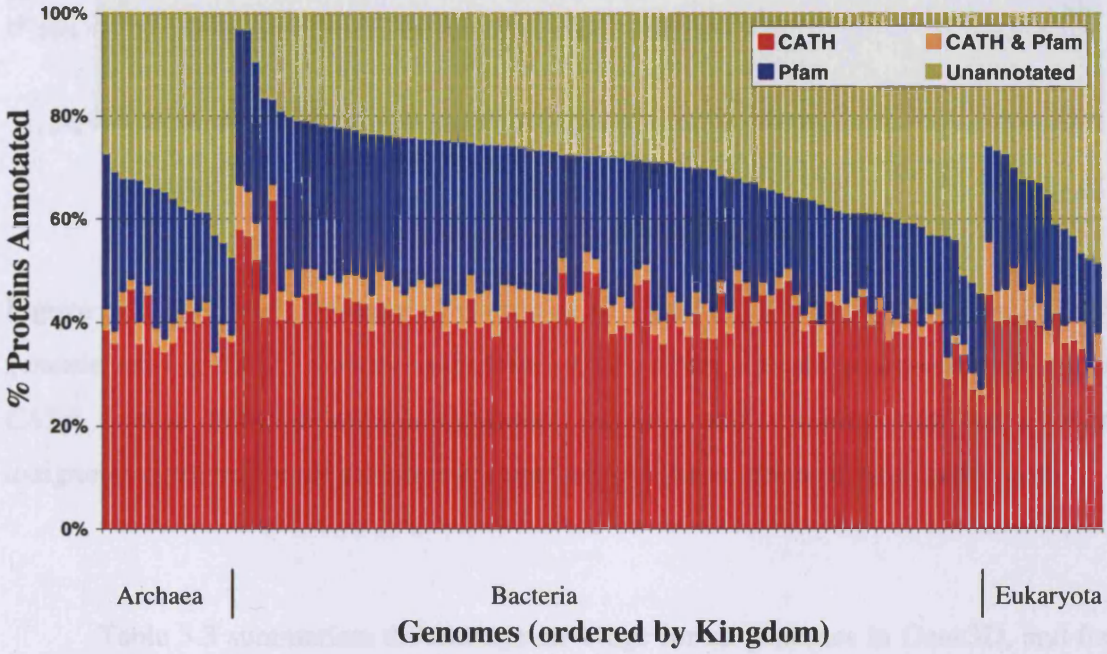


**Figure 3.12 Gene Coverage of Genomes in Gene3D.** *Percent of genes in each genome with a CATH domain assignment (red), Pfam domain assignment (blue), a CATH and a Pfam domain assignments (orange) and genes with no domain assignment (green) for all genomes grouped into Archaea, Bacteria and Eukaryota.*

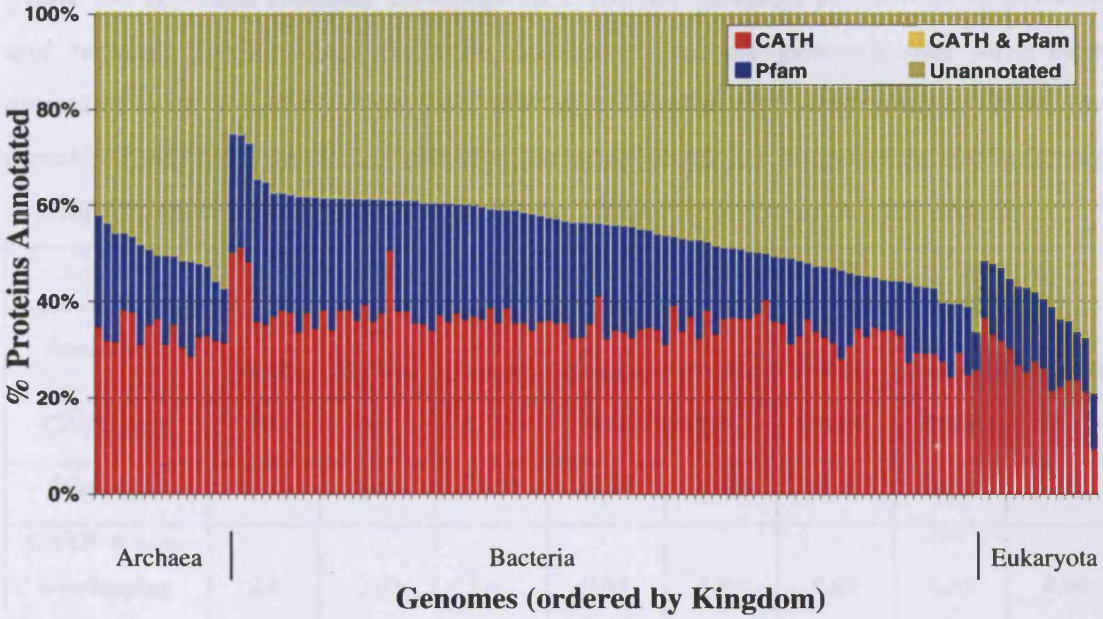


**Figure 3.13 Residue Coverage of Genomes in Gene3D.** *Percent of residues in each genome with a CATH domain assignment (red), Pfam domain assignment (blue), a CATH and a Pfam domain assignments (yellow) and residues with no domain assignment (green) for all genomes grouped into Archaea, Bacteria and Eukaryota.*

Table 3.3 summarises the average coverage for all genomes in Gene3D, and for each kingdom in Gene3D.

**Table 3.3 Average Domain Coverage in Gene3D.** *Average percentage of proteins and residues for Archaea, Bacteria, Eukaryota and all genomes that have been annotated with a CATH, Pfam, a CATH & a Pfam domain assignment. Total un-annotated and total annotated with domain assignments for all genomes are indicated in bold. Numbers in parentheses indicate the number of genomes in each dataset.*

| | **Archaea (16)** | | **Bacteria (90)** | | **Eukaryota (14)** | | **All (120)** | |
|---|---|---|---|---|---|---|---|---|
| Annotation: | % ***Protein*** | % ***Residue*** | % ***Protein*** | % ***Residue*** | % ***Protein*** | % ***Residue*** | % ***Protein*** | % ***Residue*** |
| **CATH only** | 38.81 | 33.42 | 41.88 | 35.49 | 38.15 | 26.04 | 41.04 | 34.11 |
| **Pfam only** | 22.17 | 16.30 | 24.33 | 19.37 | 19.72 | 13.80 | 23.50 | 18.31 |
| **CATH & non-overlapping Pfam** | 2.8 | 0.03 | 4.41 | 0.05 | 5.81 | 0.03 | 4.36 | 0.04 |
| **Total Unannotated** | 36.22 | 49.44 | 29.38 | 45.09 | 36.32 | 60.14 | 31.10 | 47.42 |
| **Total Annotated** | 63.78 | 50.56 | 70.62 | 54.91 | 63.68 | 39.86 | **68.90** | **52.58** |

Combination of CATH and Pfam annotations gives an average overall coverage of 68.9% of proteins within a genome having at least one domain assignment. In some small bacterial genomes (for example *Buchnera*) the total coverage is as high as 97%. On average 45.4% of proteins in a genome can be assigned at least one CATH structural domain. In addition, Pfam domain assignments can be made to a further 23.5% of proteins. An average of 4.36% of proteins in a genome comprises non-overlapping domains from both CATH and Pfam. It should be noted that genome coverage by proteins with an assignment is on average 16% higher than genome coverage by residues, indicating that whilst many proteins have a domain assignment, these domain assignments do not characterise all the residues in these proteins and so do not fully describe their domain architecture.

Genome coverage in Gene3D is comparable to the coverage in related databases, see table 3.4.

**Table 3.4 Genome Coverage in Gene3D.** *Gene coverage and residue coverage for Gene3D and comparable databases.*

| Database | Release | Gene Coverage (%) | Residue Coverage (%) |
|---|---|---|---|
| Gene3D (CATH) | Aug 2003 | **45.4** | **34.2** |
| Gene3D (CATH & Pfam) | Aug 2003 | **68.9** | **52.6** |
| SUPERFAMILY | Feb 2005 | **57.4** | **49.5** |
| Genomic Threading Database | Feb 2005 | **81.6** | **61.6** |

A lower Gene3D CATH coverage compared to SUPERFAMILY and the Genomic Threading Database is because the CATH fold library is smaller than the SCOP fold library used to provide SUPERFAMILY and The Genomic Threading Database domain assignments and more stringent criteria are used in the match/query overlap between the HMM and the protein sequence during domain assignment.

### 3.3.7 Increasing Genome Coverage in Gene3D

#### 3.3.7.1 HMM Library Expansion

Gene3D CATH domain assignment HMM libraries are built using CATH s35 sequence family seed sequences provided from the CATH database. However, seed sequences from additional sequence families in CATH may provide a mechanism to capture more sequence diversity within CATH domain families. HMMs were built as described previously, but using seed sequences from CATH s95 sequence families to produce an s95 CATH HMM library. This library was used to scan the *Escherichia coli K12* genome in order to compare genome coverage with that previously achieved using the s35 CATH HMM library, shown in table 3.5 below.

**Table 3.5 Escherichia coli Genome Coverage using s35 and s95 HMM Libraries.**
*Gene coverage (% of E.coli K12 genes with a CATH domain assignment).*

| HMM Library | Scan | Library Size | Gene Coverage (%) |
|---|---|---|---|
| s35 CATH HMMs | E.coli K12 | 4023 models | 52.36 |
| s95 CATH HMMs | E.coliK12 | 7913 models | 54.04 |

As can be seen from the table, an increase in genome coverage of 1.68% was achieved using the s95 CATH HMM library. However, this slight increase in genome coverage required a doubling in the size of the HMM library. The computational time required for scanning the genome is also doubled. This small increase in genome coverage is simply not worth the huge increase in computational resources which it takes to achieve. Similar results were obtained by Sillitoe *et al.* (Sillitoe *et al.*, 2005) where the authors used s35 and s95 CATH HMM libraries to scan a dataset of 4,036 sequence homologues, none of which had more than 35% or 95% sequence identity respectively to the HMM being scanned against. The authors concluded that coverage of their dataset was not increased significantly by using an s95 CATH HMM library over an s35 CATH HMM library.

3.3.7.2 Updated Versions of CATH and Pfam

Updated releases of both the CATH and Pfam databases contain more domains in established domain families as well as novel domains in novel domain families, and thus cover a larger amount of sequence and structure space. In addition, HMMs built using more recent and larger non-redundant protein sequence databases (see methods) are more sensitive (Sillitoe *et al.*, 2005). Successive releases should therefore provide an increase in genome coverage. All the sequences in Gene3D were scanned with CATH version 2.4 and CATH version 2.5 s35 HMMs, as well as Pfam version 10 and Pfam version 13 HMMs. In addition, SAMOSA HMMs (models built by Ian Sillitoe) were scanned against all sequences in Gene3D. SAMOSA models are built from multiple structural alignments of large CATH domain families and have previously been shown to increase coverage in specific datasets (Sillitoe *et al.*, 2005). An increase in genome coverage of 6-7% was achieved using these expanded HMM libraries (see table 3.6).

**Table 3.6 Genome Coverage in Gene3D using various HMM Libraries.** *Gene coverage (average % of genes with a domain assignment) for all genomes in Gene3D.*

| HMM Library | Release | Date | Library Size | Gene Coverage (%) |
|---|---|---|---|---|
| CATH | 2.4 | 02/02 | 3,285 | 38.1 |
| CATH | 2.5 | 08/03 | 4,036 | 45.4 |
| Pfam | 10 | 07/03 | 6,190 | 61.2 |
| Pfam | 13 | 04/04 | 7,426 | 67.8 |
| CATH + SAMOSA | 2.5 | - | 4,725 | 45.5 |
| CATH + Pfam | 2.5 / 10 | - | 10,226 | 68.9 |
| CATH + Pfam | 2.5 / 13 | - | 11,462 | 72.9 |

The insignificant increase in genome coverage found when using SAMOSA HMMs underlines the difference between genomic datasets and those previously used to benchmark SAMOSA models which showed an increased coverage of ~10% (Sillitoe *et al.*, 2005).

The trends seen above in genomic datasets are supported by the observations of Sillitoe *et al.*, who illustrate how the increase in the size of Genbank sequence repository and the increased number of CATH structural families is responsible for the increased rate of detection of remote homologues from the CATH structural database over time. The author's note that as Genbank increased in size from 907,000 sequences used for CATH version 2.4 HMM building to 1,399,000 sequences used for CATH version 2.5.1 HMM building, an increased coverage of 6% was observed in their dataset consisting of 4,036 remote homologues with less than 35% sequence identity to the models being matched. Future releases of both CATH and Pfam are likely to provide increased coverage not only due to an increase in the domains described by these resources but also as future Genbank sequence databases will be significantly larger and produce more sensitive HMM libraries.

### 3.3.8 Kingdom Distribution of Protein Families and Domain Families in Gene3D

It is interesting to consider the proportion of domain or protein families common to all the genomes in Gene3D and the percentage of genome sequences belonging to these families. For a domain to be considered universal it should be detected in a significant proportion of organisms within a kingdom, in order to reduce the likelihood of its presence being due to recent horizontal transfer between organisms as opposed to vertical transfer from a common ancestor.

However, universal domains will not necessarily be present or detected in all 120 genomes in Gene3D for several reasons. Some of the organisms are obligate intracellular parasites (*e.g. Chlamydia trachomatis*) and symbionts (*e.g. Wigglesworthia brevipalpis*) and cannot independently perform some of the functions that are essential to life (Zomorodipour and Andersson, 1999). The domain assignments in these genomes are unlikely to reflect their evolutionary heritage since protein functions and their associated domains may no longer be present in these genomes. Furthermore, HMMs used for domain mapping are not expected to identify all remote homologues. Indeed, as previously described, the level of domain assignment in genomic datasets has not saturated, and as sequence databases and domain classifications become larger more domain assignments will be made to genomes. Currently 76-80% of remote homologues are detected using the CATH HMM library (see Chapter 2, section 2.4.2.2).

Taking these factors into account we decided that domain families found in at least 70% of the organisms could be considered universal. Using this 70% universality measure the percentage of CATH and Pfam domain assignments in each genome that are probably universal to each Kingdom can be calculated. Figure 3.14 shows the percentage of CATH and Pfam domain assignments within a genome where the domain is universal in one, two or three kingdoms. Domain assignments unique to the genome are also indicated. Where a domain occurs in more than one genome but does not occur in 70% of genomes within any Kingdom and cannot therefore be considered universal to any Kingdom, the domain is indicated as not specific to a Kingdom (i.e. it is specific to only a subset of a Kingdom).

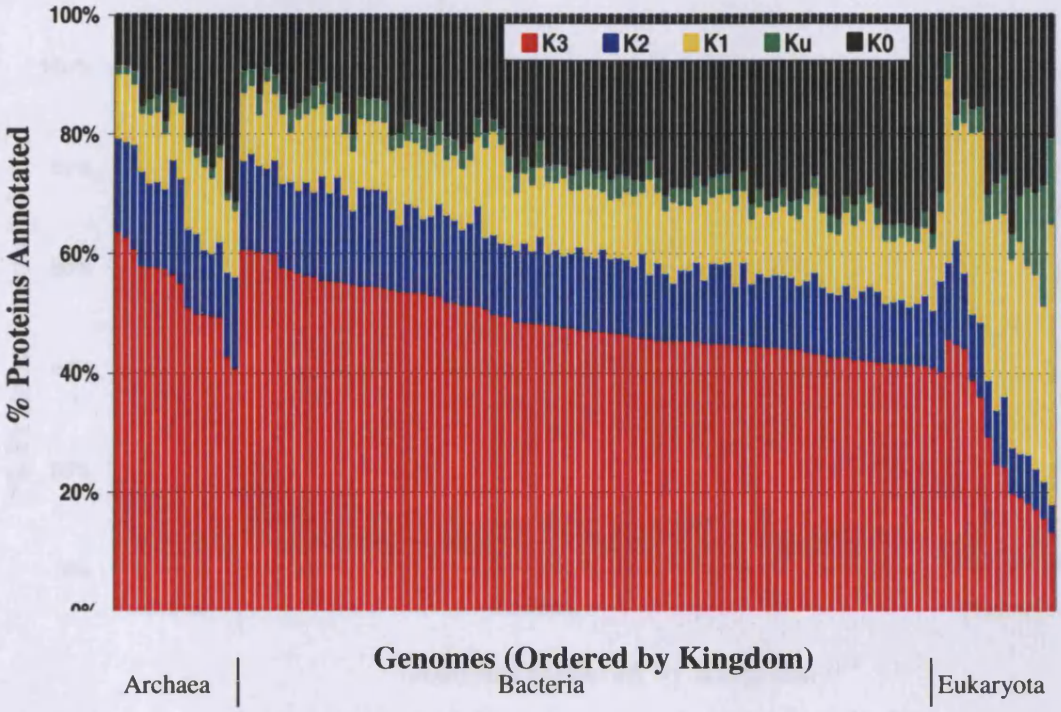


**Figure 3.14 Kingdom Distribution of Domains in Gene3D.** *Percent of proteins in each genome belonging to a domain family that is universal to one (K1 - yellow), two (K2 - blue), three (K3 - red) Kingdoms, unique to the genome (Ku - green) or specific to a subset of a Kingdom (K0 - black).*

It can be seen from figure 3.14 that ~16% of domain assignments within a genome belong to domain families which are not universal to any kingdom, while almost 50% of all domain assignments are universal to all three kingdoms of life. There are 212 CATH and Pfam domain families that are found in at least 70% of the genomes from each of the three kingdoms of life and these domains may correspond to universal families with essential functions (listed in appendix II). In contrast, figure 3.15 shows that less than 10% of protein sequences (mostly comprising multidomain architectures) are assigned to protein families universal to all three kingdoms of life while ~63% do not appear to be universal to any kingdom, as they occur in less than 70% of the organisms in any of the kingdoms.

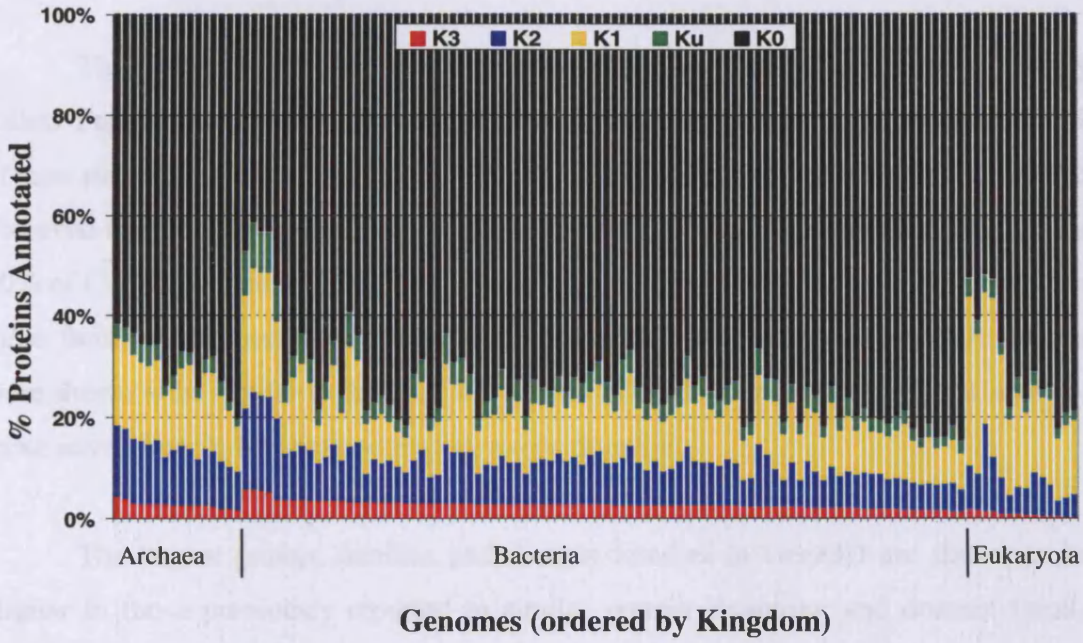


**Figure 3.15 Kingdom Distribution of Protein Families in Gene3D.** *Percent of proteins in each genome belonging to a protein family that has members from one (K1 - yellow), two (K2 - blue), three (K3 - red) Kingdoms, unique to the genome (Ku - green) or specific to a subset of a Kingdom (K0 - black).*

Interestingly, 50% of the domain structure annotations arise from families that are common to all the genomes. By contrast less than 10% of genome sequences are assigned to protein families common to all kingdoms. This is in agreement with previous findings (Chothia *et al.* 2003; Hegyi *et al.* 2002) suggesting that common domains have been combined in different ways to generate kingdom specific domain architectures. Since modification in domain architecture is frequently associated with change in protein function (Todd *et al.*, 2001), changes in domain architecture provide a mechanism for expanding the functional repertoire of the organism. These findings will be discussed and expanded later, but show that Gene3D protein families and domain assignments can be used to gain insight into the evolutionary relationships between genomes, protein families and protein domains.

## 3.4 Summary

The family size distribution of protein families in Gene3D was found to closely follow Pareto's 80/20 Law, where 20% of non-singleton protein families contain 70% of non-singleton protein sequences. Power law family size distributions were also observed in domain assignment data, where 20% of CATH domain families account for 80% of CATH domain assignments. The size distributions of Newfam regions indicate these families are mostly extremely small. Length distributions of Newfam regions were shown to be similar to those of individual CATH domains, indicating that many of these novel domain-like regions may be single domains.

The largest protein families and domain families in Gene3D are shown to be similar to those previously reported in similar protein clustering and domain family classifications. In addition, the most sequence diverse and sequence invariant protein families were functionally characterised, revealing that highly invariant protein families perform very specific functions, interacting with few ligands, whilst highly diverse protein families perform many more functions and interact with a huge variety of different ligands.

The domain architectures identified in Gene3D protein families were shown to be highly conserved within protein families, 87% of the protein families were found to have more than 70% of members sharing a common domain architecture. The power law distribution of domain architecture superfamilies across protein families was described, highlighting the potential for development of HMMs representing specific domain architectures in fast assignment of newly completed genomes.

Analysis of Genscan predicted protein sequences in Gene3D, showed that up to 57% of eukaryotic Genscan predictions are likely to be correct protein sequence predictions.

The distribution of domain families and protein families across the genomes in Gene3D revealed that whilst 50% of domain assignments in a genome are common to all three Kingdoms of life, less than 10% of proteins in a genome are assigned to protein families common to all three Kingdoms, indicating that common domain families are reused in different contexts providing diverse protein functions across the genomes.

# CHAPTER FOUR

# Application of Gene3D to Structural Genomics

## 4.1 Introduction

Many structural genomics initiatives are currently in progress and although the aims vary between consortia, many groups are selectively targeting protein families for which the fold is unknown in order to increase our knowledge of fold space.

In order to attempt predictions on fold space we need to know how many protein families there are in nature and how many of these are likely to possess a novel fold. Genome sequencing still considerably outpaces the structure genomics initiatives with more than 260 completely sequenced genomes, yielding over a million protein sequences at the start of 2005 (GOLD database, Bernal *et al.*, 2001). This contrasts with 30,041 PDB entries (Deshpande *et al.*, 2005), some 1000 of which were determined by structure genomics consortia over the last five years. Encouragingly, and in parallel with the expansions in the structure and sequence databanks over the last decade, HMM based sequence homology detection methods allow the granularity of protein family space to be more accurately charted by allowing recognition of extremely distant homologues.

Although analyses of completed genomes suggest that there are tens of thousands of domain families (Lee *et al.*, 2005; Liu and Rost, 2002), currently only 5% of newly determined structures are observed to have a novel fold (Todd *et al.*, 2005), suggesting the there are a much smaller number of folds in nature. The Protein Structure Initiative (PSI; www.nigms.nih.gov/Initiatives/PSI/) is a large-scale, high-throughput structural genomics project with groups from the United States, European Union, Japan, China, Canada and Israel. The PSI entered a production phase, with the aim of solving structures for all the large structurally uncharacterised domain families, to increase the number of known folds in the PDB. Around 2% of structures deposited in the PDB by conventional structural biology contain novel folds. In contrast, 11% of structures deposited in the PDB during the initial phase of the PSI (September 2000

until June 2005) were novel folds, and 67% were structures for new sequence families, compared to 21% from conventional structural biology (Todd *et al.*, 2005).

Thus it is clear from this initiative that even when sequence families are targeted because they are predicted to have a novel fold, a relatively small percent (~11%) are found to possess a novel fold once their structures are solved. These observations support earlier hypotheses (Chothia, 1992; Orengo, 1994) derived from analyses of sequence data that there are a limited number of folds in nature.

Over the last decade there have been several attempts to predict the number of folds. Whilst Wolf *et al.* (2000) predict the number of folds in individual genomes; most estimates consider the total number of folds in nature. Current estimates of the number of folds range from 1000 to 10,000 depending on the models and approximations applied (Leonov *et al.*, 2003; Coulson and Moult, 2002; Koonin *et al.*, 2002).

One of the earliest estimates of fold numbers was a simple approximation by Chothia (1992). This assumed that there are a limited number of folds in nature that sequences could adopt due to physical constraints. If these are randomly sampled then the probability that a new sequence has a novel fold can be estimated by determining the proportion of unrelated sequences e.g. in the structure classification SCOP that are found to share the same fold. This approach predicted 1000 structural families based on the proportion of sequences of known structure in SCOP that had unique folds, the fraction of the SWISSPROT sequence database these sequences comprised and the fraction of new sequences found to be related to sequences in SWISSPROT. A similar model applied by Orengo and co-workers (1994) also took account of the number of protein families in SWISSPROT. Using the CATH structure database they predicted a higher estimate of ~8000 folds. Both these calculations were undertaken when sequence and structure databases were relatively sparse, and under-represent the bias in the distribution of certain folds, often referred to as superfolds (Orengo *et al.*, 1994), which are more highly reused by different protein families in nature than expected by chance.

Since Chothia's early estimates, several groups have applied approaches that model this uneven distribution of fold usage (Zhang and DeLisi, 1998; Govindarajan

and Goldstein, 1996). Random sampling of known sequence families assigning equal likelihood to each fold gives rise to a non-uniform fold distribution which, when further modified to account for the extreme bias of the superfolds and the fact that many folds are only rarely seen in nature, gives an estimate of 4000 folds (Govindarajan *et al.*, 1999).

Coulson and Moult (2002) assume three types of folds – superfolds which are adopted by very many protein families and are highly recurrent in the genomes, mesofolds which have an intermediate number of protein families associated with them and unifolds adopted by a single narrow sequence family. They simulated the expansion of new folds classified in the SCOP structure database over the last two years, as a fraction of new sequence families added. Assuming a maximum of 50,000 protein families in nature, this approach predicts up to 400 mesofolds and some 10,000 unifolds in addition to 9 superfolds. Perhaps more importantly, the majority of sequence families belong to superfold and mesofold groups and for 80% of these families we probably know the fold already.

Several groups attempt to model the uneven fold/family distribution using power laws. Power laws appear to be ubiquitous in nature and society and seem to explain many of the biological trends recently revealed by genome data e.g. protein family distributions, domain associations, protein-protein interactions (Koonin *et al.*, 2002; Qian *et al.*, 2001; Luscombe *et al.*, 2002).

Karev *et al.* (2003) model protein family distributions by simulating the birth (gene duplication), death (gene loss) and innovation (new protein) of different domains in individual genomes (Karev *et al.*, 2002). Although this entirely stochastic model fails to account completely for the observed distribution, it shows that a close fit is possible using a model with only three independent parameters. Implicit in the model is the notion that the 'fit' get 'fitter' and domains randomly duplicated early in evolution increasingly dominate the population. None of these models incorporate selection pressures that might operate to favour the retention of duplicated domains performing important biochemical activities. However, many highly recurrent domains appear to have important biochemical functions; for example in providing energy or redox equivalents for enzyme reactions or in responding to cellular signals and binding to DNA (Pawlowski *et al.*, 2001).

These models still ignore possible bias in the structure and sequence databases. However, it is likely that proteins sampled for structure determination have been relatively easy to purify and crystallise - witnessed by the small numbers of transmembrane structures known. Perhaps more worrying are recent analyses suggesting that we have barely sampled sequence and family space as each new genome adds more families and there is no sign of saturation in this expansion (Kunin *et al.*, 2003). Even with the huge advances in genome sequencing, there are still at least ten million more organisms uncharacterised (Koonin *et al.*, 2002).

### 4.1.1 How many Domain Families are Currently Recognised and how many Novel Folds can we predict using this data?

One of the first steps in calculating how many new folds remain to be discovered is the determination of the number of sequence families in nature. Once we have a reasonable prediction for this number we can estimate the number of new folds based on the proportion of structurally characterised families that have unique folds. Perhaps the hardest problem in clustering sequences into protein families is handling the similarities between multi-domain proteins and the fact that many different multi-domain proteins share common domains but in different contexts. This recurrence of domains suggests their importance as primary evolutionary units and although some researchers hypothesise that smaller super-secondary structural motifs may be the building blocks of evolution (Soding and Lupas, 2003), the majority of globular compact folds characterised to date comprise whole domains.

However, domain boundary recognition is a non-trivial algorithmic challenge particularly if no structural data are available. Even methods based on structures disagree in their assignments 20-40% of the time (Jones *et al.*, 1998). The problem is compounded by discontinuities in some domain sequences whereby the insertion of a second domain disrupts an existing domain region within a multidomain protein. Structural data in CATH suggests these discontinuities exist in about 23% of domains occurring in multi-domain proteins (Pearl *et al.*, 2002). In Gene3D, there are 341,726 domain assignments made to multi-domain proteins, of which 31,090 or 10.0% are discontinuous domain assignments.

Some of the most successful approaches to boundary prediction combine multiple sequence data and residue propensities using neural networks (Liu and Rost, 2003; Yona and Levitt, 2000). Other methods exploit the recurrence of domains in different contexts to identify boundaries from multiple alignments (Servant *et al.*, 2002; Park and Teichmann, 1998; Heger and Holm, 2003). The elegant approach of Holm and co-workers (ADDA) exploits graph theory to build networks of domain links in multidomain proteins from which multiple alignments can be extracted and recursively analysed and chopped to yield their single domain components.

Estimates of the numbers of domain families identified vary substantially depending on the sequence datasets clustered and thresholds employed. The ADDA algorithm of Holm and co-workers which firsts chops sequences into domains and then clusters, identifies some 34,000 domain families in a combined sequence dataset of SWISSPROT, TrEMBL, PIR, PDB, WORMPEP and ENSEMBL which after removing redundancy at 40% sequence identity, contained almost 250,000 protein sequences. Almost 170,000 sequences remain as singleton sequences that are not clustered into any family.

Similarly, a recent analysis by Liu and Rost (2002), chopping and clustering sequences from 5 eukaryotic genomes suggested 22,000 domain-like clusters in eukaryotes. Again these represent low estimates as only a tiny percentage of species have been completely sequenced. Additional analysis by Liu and Rost (2002), chopping and clustering sequences from 62 complete genomes identified 118,108 singleton and 63,300 non-singleton domain-like clusters.

Although, similarity in the structures adopted by different families may reflect folding preferences and convergence to energetically stable folds it is possible that many of the families adopting the superfolds are in fact very distantly related, beyond the sensitivity of current algorithms to detect homology. Families adopting TIM barrel folds are a case in point with recent analysis suggesting that many families may have evolutionary links supported by unusual sequence signatures and functional properties (Copley and Bork, 2000; Nagano *et al.*, 2002).

Recent calculations of the number of folds using new estimates of sequence families, suggest about 400-6000 well populated folds in nature (Grant *et al.*, 2004; see

also section 4.3.3 below). The PSI proposes to solve 3000 structures within the next five years (Chandonia and Brenner, 2005). Thus, provided families are targeted carefully, we may know a significant proportion of all highly populated folds in nature by the end of the initiative. However, due to the very high attrition rates, target selection is important for increasing the efficiency of future structural genomics initiatives. In the first phase of the PSI, which ran from September 2000 until June 2005, only 2-10% of the proteins selected by PSI structural genomics centres resulted in a solved structure. For example the Midwest Center for Structural Genomics Center targets around 5000 proteins a year from which only 100-200 are solved (www.mcsg.anl.gov/). The loss of targets at each stage in the process of structural determination is shown in figure 4.0 below.

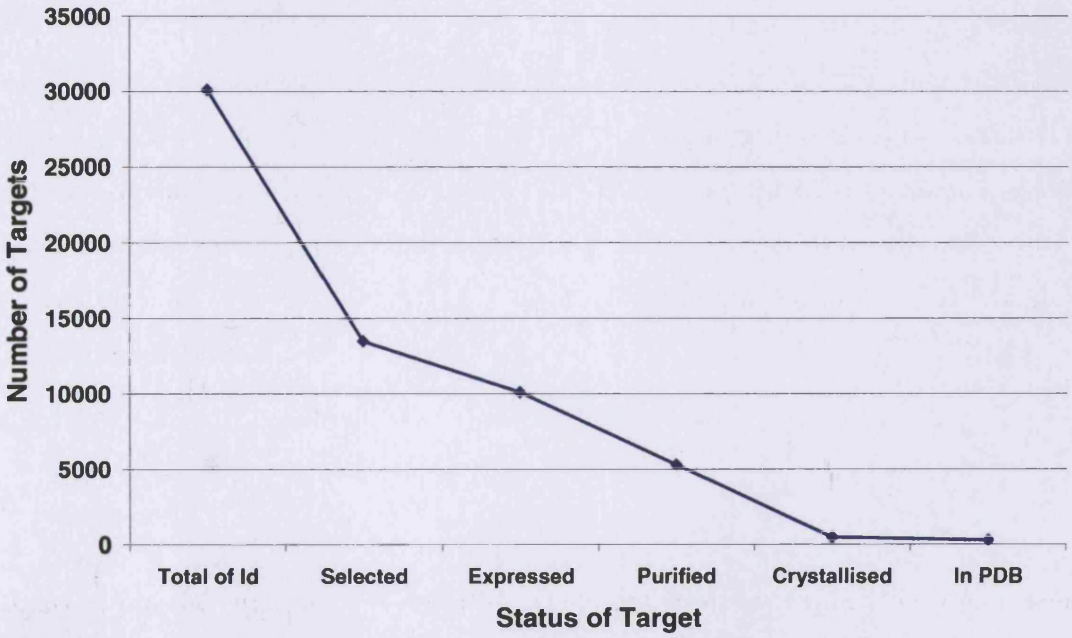


**Figure 4.0 The Number of Targets at each stage of the Structural Genomics Pipeline.** *From over 30,000 elected targets, 314 structures were deposited in the PDB in 2003. Adapted from Bourne et al., 2003.*

We now know the fold for many of the largest domain families, particularly those which dominate the genome annotations, for example 813 CATH folds can be assigned to 45.4% of protein sequences in the genomes, and an additional 4,440 Pfam families can be assigned to approximately 23.5% of protein sequences (see figure 3.3.6 and table 3.4). Once a fold has been identified for one member of these domain

families, it can be confidently predicted that all other domain family members will adopt a similar fold. However, it may not be possible to model every member of the domain family to give accurate models, since many relatives are likely to have a prohibitively low sequence identity to the relative with solved structure. Vitkup *et al.* (2001) suggest that 30-35% sequence identity is sufficient to achieve accurate comparative modelling, and recent analysis of CATH supports this view (Reeves *et al.*, in preparation). Therefore in addition to targeting structurally uncharacterised sequences that have been predicted to have a novel fold, it is clear that in some families multiple targets must be solved to increase the number of accurate models available for the family.

## 4.2 Objectives

This chapter describes using Gene3D for selection and prioritisation of targets for structural genomics initiatives. The largest structurally uncharacterised domain families occurring in the genomes are identified as primary targets, solving at least one structure for these families will increase the proportion of genome sequences with a known fold. This has been referred to as 'coarse grained' structural coverage of genome sequences. Additional targets are then identified in large families which are already structurally characterised but are under-represented in the PDB. Solving these structures increases the number of sequences in these families for which accurate homology models can be built, so called 'fine grained' structural coverage of genome sequences. A protocol for prioritising these structural genomics targets will be shown, that allows rational target selection.

## 4.3 Results

### 4.3.1 Calculating the Number of Domain Families in Gene3D

The following section details the identification of sequence families in Gene3D data, including CATH domain sequence families of known structure, Pfam domain sequence families, and Newfam domain sequence families. Newfams are novel domain families identified in Gene3D (described previously, see section 2.4.2.8). Analysis of the distribution of these sequence families can suggest the total number of fold groups present in the 120 genomes in Gene3D and can be used to identify families suitable for coarse grained target selection i.e. to increase the number of known folds and structural families.

Sequence families for CATH domains (CATH-fams), non-overlapping Pfam domains (Pfam-fams) and novel domains containing neither CATH nor Pfam assignments (Newfams) were characterised from s35 protein family subclusters in Gene3D (see section 2.3.3).

50% of Gene3D domain assignments to genomes can be assigned to 93,571 CATH-fams representing 1277 CATH homologous superfamilies from 813 folds. A further 33% of domain assignments can be assigned to 61,722 Pfam-fams representing 1832 Pfam domain families (Pfam release 9). The remaining 17% of domain assignments are assigned to 52,973 Newfams. The number of sequence families of each family type (CATH-fam, Pfam-fam, Newfam) is shown in table 4.0 below. As can be seen from table 4.0, there are a total of 208,266 sequence families in Gene3D.

**Table 4.0 Domain Family Characterisation in Gene3D.** *This table excludes 148,578 singleton sequence families not considered in this analysis.*

| Family Type | Percentage of Total Domains | Number of s35 Sequence Families | Number of Superfamilies | Number of Folds |
|---|---|---|---|---|
| CATH-fam | 50 | 93,571 | 1277 | 813 |
| Pfam-fam | 33 | 61,722 | 1832 | - |
| Newfam | 17 | 52,973 | - | - |

The uneven fold/family distribution, revealed by several previous analyses (Zhang and DeLisi, 1998; Govindarajan and Goldstein, 1996) can be clearly seen in figure 4.1 below, which shows that a small percentage of fold groups in the CATH domain structure database - 54 'superfolds', (defined as CATH folds with three or more homologous superfamilies), representing only 6.6% of the number of fold groups in CATH, are very highly populated, accounting for 76% of s35 sequence families, whilst there are many folds adopted by a single family.

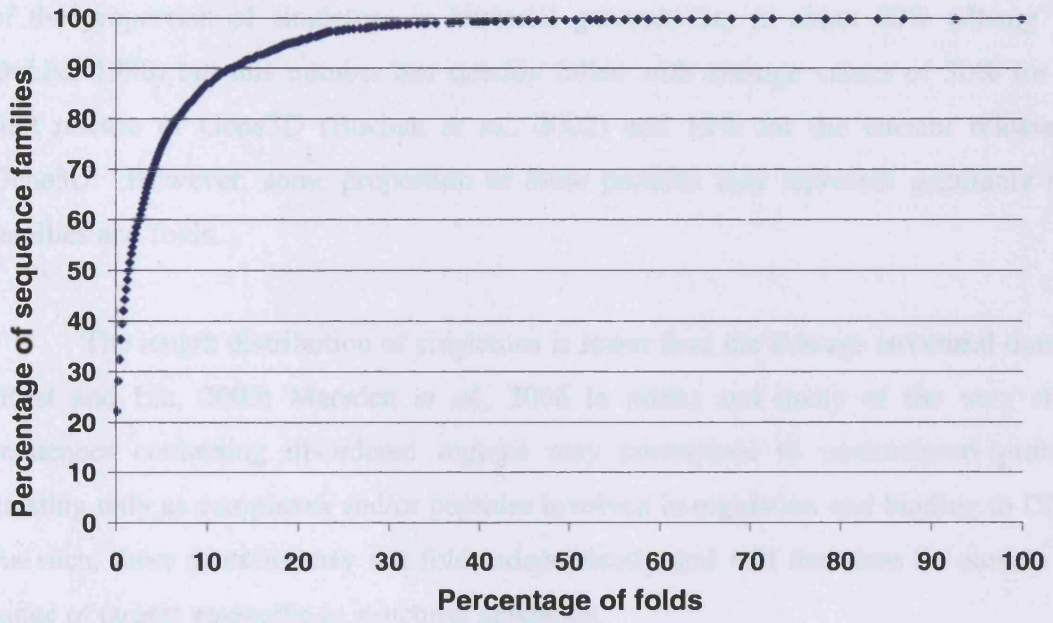


**Figure 4.1 Percentage of CATH Folds Accounting for Percentage of CATH s35 Sequence Families in Gene3D.** *The percentage of CATH folds against the cumulative percentage of CATH sequence families shows marked sequence family size distribution bias where a small number of folds account for a large proportion of sequence families.*

Although, similarity in the folds adopted by different families may reflect folding preferences and convergence to energetically stable folds it is possible that many of the families adopting the superfolds are in fact very distantly related, beyond the sensitivity of current algorithms to detect homology. Families adopting TIM barrel folds are a case in point with recent analysis suggesting that many families may have evolutionary links supported by unusual sequence signatures and functional properties (Copley and Bork, 2000; Nagano *et al.*, 2002).

### 4.3.2 Should Structural Genomics be Targeting Singletons?

Are the singleton sequence families (almost 150,000 in Gene3D) distant relatives of existing families unrecognised by current algorithms or are they genuinely unique sequences having novel folds, and therefore good targets for structural genomics? Kunin and co-workers (2003) showed that as newly sequenced genomes are completed, there is a constant gain in the number of singleton families. This may change as the databases expand and recognition methods improve. Original estimates of the proportion of singletons in bacterial genomes lay at about 50% (Zhang and DeLisi, 1998) but this number has steadily fallen with average values of 30% for the first release of Gene3D (Buchan *et al.*, 2002) and 18% for the current release of Gene3D. However, some proportion of these proteins may represent genuinely new families and folds.

The length distribution of singletons is lower than the average structural domain (Rost and Liu, 2003; Marsden *et al.*, 2006 in press) and many of the very small sequences containing disordered regions may correspond to unstructured proteins existing only as complexes and/or peptides involved in regulation and binding to DNA. As such, these proteins may not fold independently and will therefore lie outside the range of targets amenable to structural genomics.

### 4.3.3 How many folds remain to be discovered by Structural Genomics?

Using the numbers of s35 sequence families identified by Gene3D we can make an approximation of the total numbers of folds in completed genomes by assuming the following: (1) We have sampled all the superfolds – defined as folds with 3 or more homologous superfamilies in CATH (i.e. 71,080 CATH-fams from the 54 highly populated CATH folds); (2) We have been able to map these folds onto all their relatives in the genome sequences and so we can remove these folds from the estimates of the remaining numbers of folds; (3) Singletons can also be removed from the estimate as they are probably very distant relatives belonging to these or other folds, that have diverged beyond the sensitivity of current recognition methods or because they are short sequences unlikely to fold independently but associated with functional complexes. Although singletons could be novel folds and as such could skew any

estimate of total number of protein folds, they do not represent a significant proportion of domains; and finally (4) We assume that non-superfolds and non-singletons have been sampled randomly by families in nature and that there are no biases in the current sequence and structure databases.

Removing the 54 superfolds from the Gene3D dataset leaves (93,571 – 71,080) 22,491 CATH-fams of known structure which adopt (813-54) 759 folds in CATH. Therefore, we can expect the 114,695 remaining s35 sequence families in Gene3D of unknown structure (61,722 Pfam-fams + 52,973 Newfams) to adopt ((114,695/22,491)*759) 3871 new folds. We must also take into account the fact that about 25% of the sequences of newly sequenced genomes can not be assigned to any protein family in Gene3D (Marsden *et al.*, 2006 in press), increasing the number of novel folds by 4/3 to 5161. Adding together the superfolds, known folds, and estimated number of new folds (54 superfolds + 759 known folds + 5161 estimated folds) we get an estimate of the number of folds in the 120 genomes in Gene3D of 5974. Although all fold estimates are unsatisfying in that they necessitate simplified models of fold usage and optimism regarding lack of bias in the databases and our sparse sampling of species space, the values predicted by current data suggest that provided families are targeted rationally in the next phase of the PSI, we may know a large proportion of the major fold groups by the end of the initiative.

### 4.3.4 Structural Genomics Target Selection Using Gene3D

#### 4.3.4.1 Coarse Grained Target Selection to Identify Novel Folds

There are 4,365 Pfam domain families (Pfam release 10) identified in Gene3D that are non-overlapping with CATH domain assignments, including singleton Pfams. Pfam families would make good targets for structural genomics initiatives since Pfam is a well validated and curated resource and these families represent some of the largest, structurally uncharacterised families in the genomes. Figure 3.5 (see section 3.3.2.1) illustrates the fact that Pfam families are much larger than Newfam families. In addition, they have been manually validated to improve domain boundary identification, an important consideration for structure determination. Of these targets, 1,876 Pfam families contain more than 20 members identified in the genomes, representing 89.3%

of total Pfam domain sequences; and 447 Pfam families contain more than 100 members identified in the genomes, representing 56.3% of total Pfam domain sequences.

Since protein structure initiatives propose to solve 3000 structures over the next five years, these Pfam families would be good targets for structural genomics initiatives that aim to identify novel folds, since these larger Pfam families would have a significant impact on genome coverage for each solved structure (shown in figure 4.2 below). Alternatively, these Pfam family targets could be prioritised according to Kingdom distribution, or individual genome occurrence, using Gene3D information, according to any specific aims of individual structural genomics efforts. For example, the Midwest Center for Structure Genomics consortium (MCSG) is particularly interested in targeting pathogens.

This approach is similar to the Pfam5000 strategy reported by Chandonia and Brenner (2005), in which the authors propose targeting the 5000 largest Pfam families. As the authors note, solving a single structure for each Pfam family could significantly increase coarse grained structural coverage of sequence space. However, if representative structures for the largest 1,876 Pfam families are solved this would then give structural assignments for, on average 66.4.0% of genome sequences (45.4% previously characterised by CATH domain assignments, and an additional 21.0% characterised by the 1,876 newly determined Pfam families; see genome coverage section 3.3.6 and table 3.4). Although aiming to solve 5000 or more structures may be outside the scope of the PSI initiative, 1,876 new structures may be achievable, and if the largest uncharacterised Pfam families are targeted, this will cover nearly 90% of Pfam domain sequences (see figure 4.2 below) and ensure that a significant proportion of domain sequences have a known structure (see figure 4.3).

Of these 1,876 Pfam families identified above, 1,369 families have been selected for structural determination by the PSI over the next two years. These were chosen because they had low probabilities of containing transmembrane regions, or other features which might make these targets harder to solve. To improve the probability of obtaining a structure for a selected family, each consortium typically identifies 5-10 relatives per family. Over the remaining three years of the PSI initiative, further sampling from these families may be needed and targeting of the next largest Pfam

families will be undertaken. Once all the Pfam families have been targeted, large Newfam families will be sampled.

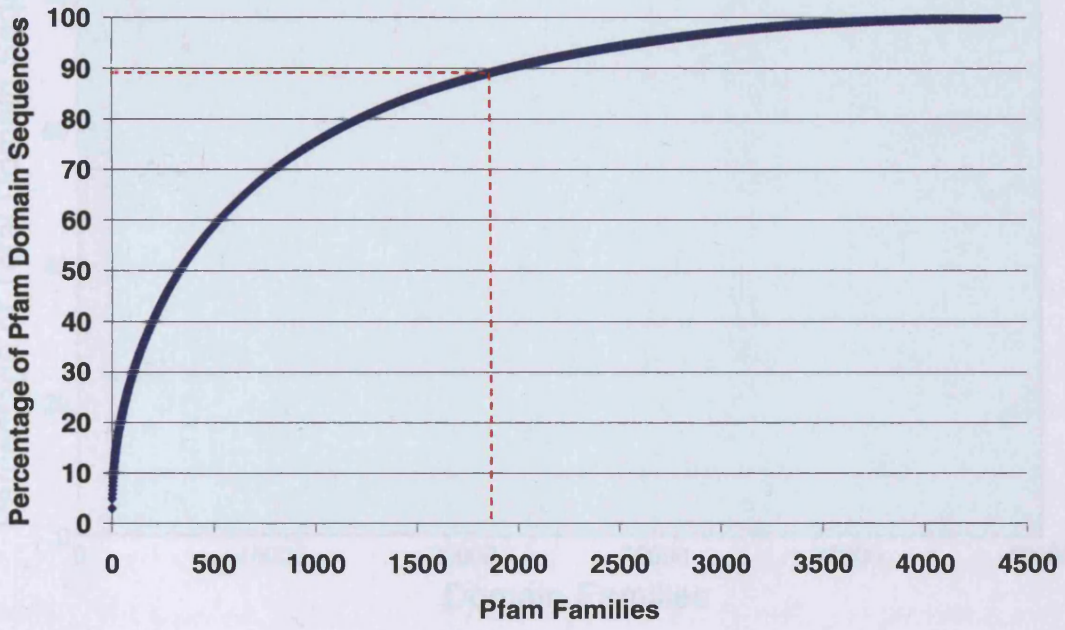


**Figure 4.2 Running Total of the Percentage of Pfam Domain Sequences in the largest Pfam families in the Genomes.** *Pfam families are ordered by number of relatives in the family. Note that the largest 1,876 Pfam families account for 89.3% of Pfam domain sequences in the genomes (red dotted lines).*

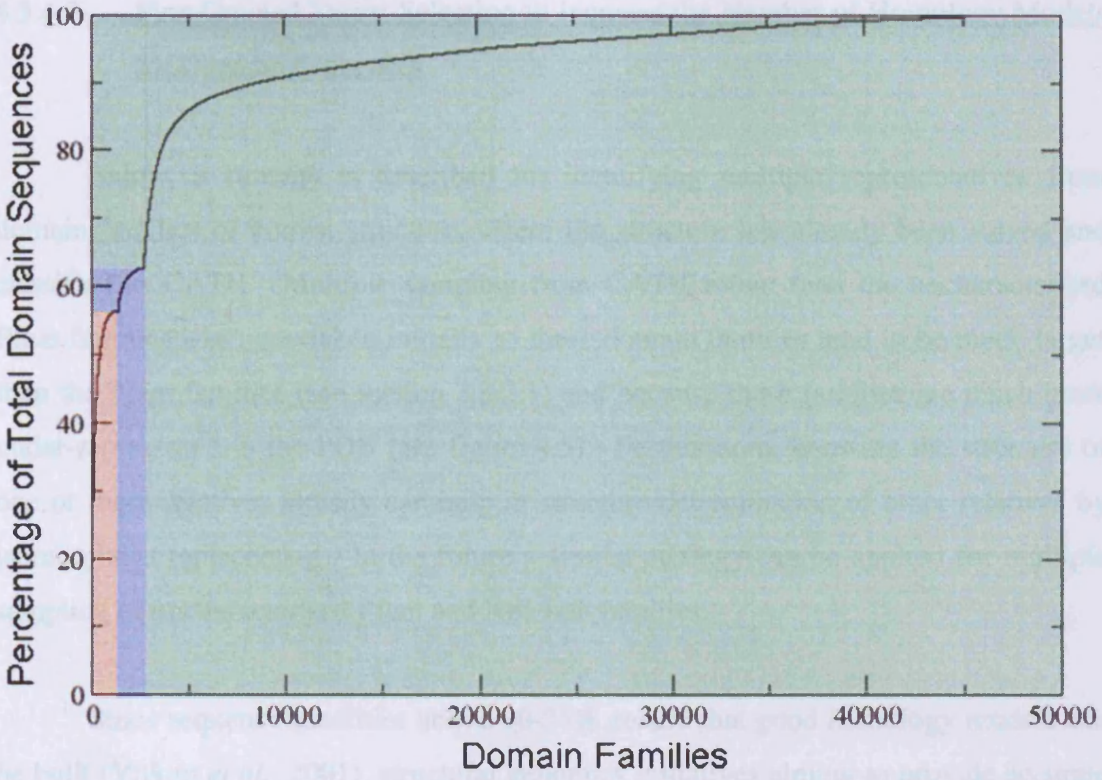


**Figure 4.3 Running Total of the Percentage of Domain Sequences in CATH, Pfam and Newfam families in the Genomes**. *CATH domain families of known structure (red), Pfam domain families of unknown structure targeted by the PSI (blue) and remaining families (untargeted Pfam families and novel Newfam families, of which ~20% are membrane associated; green), ordered by decreasing number of members plotted against the percentage of domain sequences in the genomes.*

Although solving the structures of the largest structurally uncharacterised Pfam families will increase the proportion of genome sequences for which we know the fold, as discussed above many of the sequences in families of known structure will not be closely enough related to the solved structure to build accurate homology models from this structure. Therefore to increase the number of sequences for which an accurate homology model can be built, it is necessary to select additional targets from families of known structure. This has been described as fine grained sampling.

#### 4.3.4.2 Fine Grained Target Selection to Increase the Number of Homology Models for Genome Sequences

Below, a strategy is described for identifying multiple representatives from domain families of known structure, where the structure has already been solved and classified in CATH. Multiple sampling from CATH rather than the uncharacterised Pfam families was undertaken initially as these domain families tend to be much larger than the Pfam families (see section 3.3.2.1) and because these families are much more under-represented in the PDB (see figure 4.5). Furthermore, knowing the structure of one or more relatives already can help in structure determination of other relatives by isomorphous replacement. In the future a similar strategy can be applied for multiple sampling of uncharacterised Pfam and Newfam families.

Since sequence identities above 30-35% ensure that good homology models can be built (Vitkup *et al.*, 2001), structural genomics initiatives aiming to provide accurate homology models for all members of a structural family require domain family subclusters to be identified. Clustering was performed as described previously in Chapter 2 (see section 2.3.3).

Sequence identities between domain sequences from each CATH homologous superfamily were calculated from an all against all BLAST (where at least 80% of the longer sequence is overlapped) and TCluster was then used to cluster each CATH domain family in Gene3D at 35 sequence identity (see section 2.3.3). CATH domain assignments made to genscan predicted protein sequences are included in this analysis as these genscan predictions are likely to represent translated reading frames.

Clustering of CATH domain family sequences in the 120 genomes of Gene3D produced more than 93,000 s35 subclusters. However, since this represents a very large number of targets and may be outside the scope of what is possible in current structural genomics initiatives, further analysis using Gene3D has been used to prioritise these targets and is discussed below.

### 4.3.5 Prioritising Sequence Diverse Domain Families

The most sequence diverse CATH domain families identified in the genomes, as measured by the number of distinct s35 subclusters, are shown in table 4.1 below. These comprise a total of 20,763 s35 subclusters representing 20.4% of the genome sequences. Many of these families are already observed to be structurally very diverse (Reeves *et al.*, in preparation), and solving additional structures in these families will give further insights into the manner by which structures and functions have evolved in diverse relatives.

**Table 4.1 Most Diverse CATH Domain Families in Gene3D.** *The CATH domain family name of the top ten most diverse CATH domain families is shown, with the total number of family members and the number of s35 subclusters identified in Gene3D.*

| Domain Family | Number of family members in Gene3D | Number of s35 subclusters in Gene3D |
|---|---|---|
| P-loop containing nucleotide triphosphate hydrolases | 31,908 | 4881 |
| Immunoglobulins | 19,290 | 2571 |
| DNA binding domain, transcription factor | 4456 | 2385 |
| NAD(P)-binding Rossmann-like Domain | 10697 | 2069 |
| "winged helix" repressor DNA binding domain | 8270 | 1710 |
| Hydrolase activity, aromatic compound metabolism | 4738 | 1633 |
| Periplasmic binding protein-like II | 5637 | 1439 |
| S-adenosylmethionine-dependent methyltransferase activity | 5674 | 1432 |
| YVTN repeat-like/Quinoprotein amine dehydrogenase | 5162 | 1369 |
| Ribonuclease Inhibitor | 4566 | 1274 |

Since known structures in CATH only represent a small proportion of CATH domain sequences in Gene3D, and since these may not have been sampled to reflect the true sequence diversity of these families in nature, the difference between the sequence diversity of these families in CATH was compared to the predicted diversity of families in the genomes. The relative diversity was calculated by dividing the percentage of total s35 subclusters in the genomes by the percentage of total s35 subclusters in CATH for each domain family. To enable this comparison, the CATH classification of domain families with known structure was clustered to identify s35 subclusters using the same protocol described previously for identification of s35 subclusters in Gene3D (see section 2.3.3). 193 domain families in CATH are from Kingdoms (i.e. viruses) or genomes not represented in Gene3D and are therefore excluded from this analysis. The relative diversity of domain families is shown in figure 4.5 below. There are 21-fold more s35 subclusters in total in the genomes than in CATH.

Figure 4.4 below shows the most diverse domain families in the genomes are comparatively more diverse than indicated by their CATH classification. Clearly it would be beneficial to have more structural representatives for these highly expanded domain families.

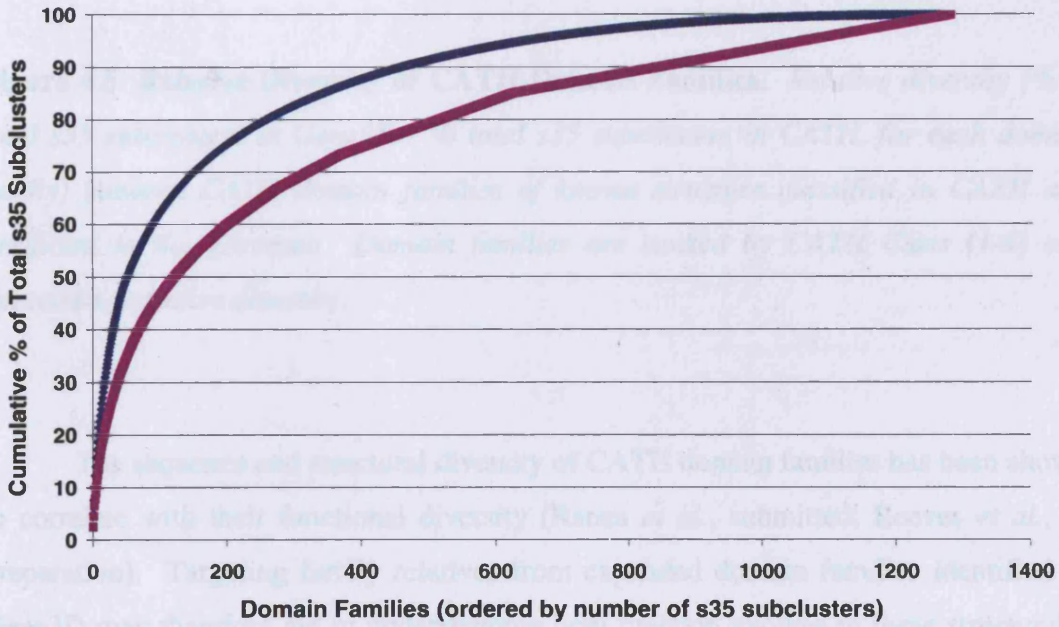


**Figure 4.4 Size/Diversity of CATH and Gene3D Domain Families.** *Domain families ranked by number of s35 subclusters (diversity) against percentage of total s35 subclusters (total diversity) for CATH (pink) and Gene3D (blue).*

As can be seen in figure 4.5 below, some domain families are much more sequence diverse in the genomes than in the CATH classification. 132 domain families with a relative diversity greater than 2 were considered significantly structurally under-represented, representing 37,214 s35 subcluster targets. Taking into account the current attrition rates of 2-10%, multiple targeting of these families could provide structures for a further 700-3000 relatives from these families.

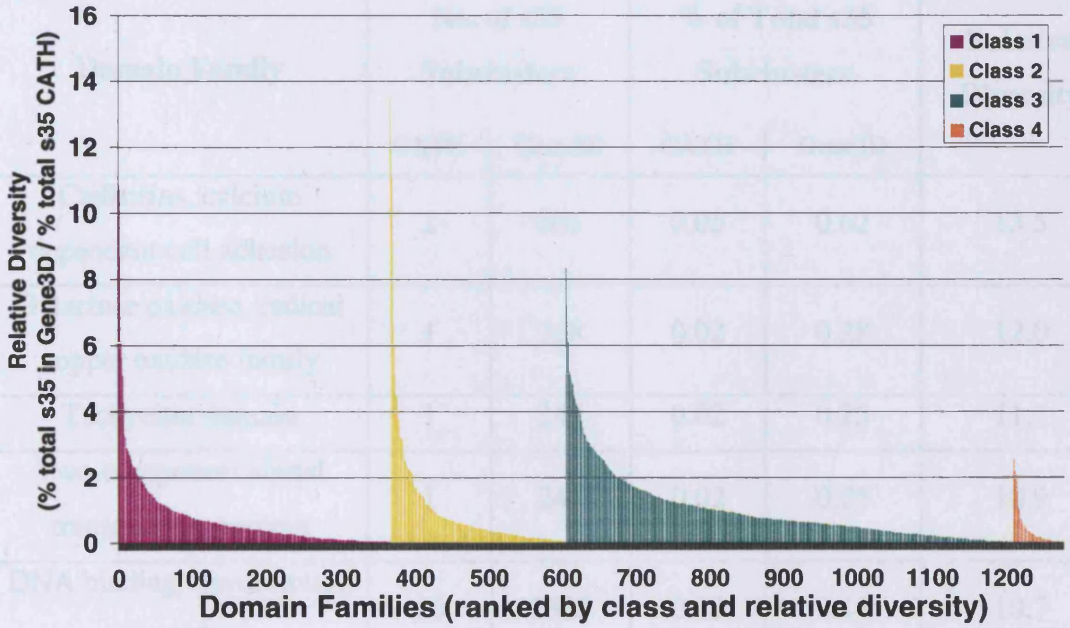


**Figure 4.5 Relative Diversity of CATH Domain Families.** *Relative diversity (% of total s35 subclusters in Gene3D / % total s35 subclusters in CATH, for each domain family) between CATH domain families of known structure classified in CATH and predicted in the genomes. Domain families are ranked by CATH Class (1-4) and descending relative diversity.*

The sequence and structural diversity of CATH domain families has been shown to correlate with their functional diversity (Ranea *et al.*, submitted; Reeves *et al.*, in preparation). Targeting family relatives from expanded domain families identified in Gene3D may therefore aid in understanding how function evolves in these structurally diverse domain families. The top most under-represented domain families (in terms of predicted structural diversity) are shown in table 4.2 below. There are 5,966 s35

subclusters altogether in these families with no structural representative, that could be targeted for structure determination.

**Table 4.2 Most Structurally Under-represented CATH Domain Families.** *The relative diversity (% of total s35 subclusters identified in Gene3D divided by the % of total s35 subclusters identified in CATH) is shown for the ten most under-represented domain families.*

| Domain Family | No. of s35 Subclusters | | % of Total s35 Subclusters | | Relative Diversity |
|---|---|---|---|---|---|
| | CATH | Gene3D | CATH | Gene3D | |
| Cadherins, calcium dependent cell adhesion | 2 | 601 | 0.05 | 0.62 | 13.5 |
| Galactose oxidase, radical copper oxidase family | 1 | 268 | 0.02 | 0.28 | 12.0 |
| Tachycitin domain | 1 | 247 | 0.02 | 0.25 | 11.1 |
| Two-component signal transduction protein | 1 | 242 | 0.02 | 0.25 | 10.9 |
| DNA binding, transcription regulation | 10 | 2385 | 0.23 | 2.46 | 10.7 |
| "winged helix" repressor DNA binding domain | 1 | 193 | 0.02 | 0.20 | 8.7 |
| Two-component signal transduction protein | 1 | 186 | 0.02 | 0.19 | 8.3 |
| Two-component signal transduction protein | 5 | 906 | 0.11 | 0.93 | 8.1 |
| AMP binding domain, peptide synthase proteins | 4 | 642 | 0.09 | 0.66 | 7.2 |
| Complement Protease C1s, immune response | 2 | 296 | 0.05 | 0.31 | 6.6 |

Some of these domain families are found in proteins performing important biological functions. Three of the domain families are involved in the two-component signal transduction pathway, an extremely functionally diverse protein family

(described previously, see section 3.3.1.2.2). Of particular interest is the galactose oxidase domain family, from a copper oxidase family of enzymes. These enzymes have a distinct active site, containing a novel Tyr-Cys modified amino acid dimer, formed spontaneously during the maturation of the protein. Interestingly, the metal ion ligands in the active site of these proteins have been shown to perform essential proton transfers and redox functions. This family of oxidases has a wide phylogenetic distribution, and may play a fundamental role in the biology of oxygen (Whittaker, 2002). The diversity across the family may reflect different interaction partners and solving the structures of multiple relatives from this family may provide insights into these interactions.

Also of functional interest is the tachycitin domain, a chitin-binding domain found in a variety of proteins. Tachycitin has been shown to have antimicrobial properties in many organisms, and in mammals in particular is thought to participate in immune defence response against nematodes and other pathogens (Tjoelker *et al.*, 2000). Finally, the complement protease C1s, the first protein of the classical component cascade system consisting of about 30 serum proteins, has a well documented biological function. Solving the structure of additional representatives may allow a greater understanding of the structural mechanisms underlying the activation, mechanism of action and substrates of relatives in this important protease family.

### 4.3.6 Prioritising Functionally Diverse of Domain Families

Another approach is to preferentially target functionally diverse and highly expanded families in the genomes, whose functional diversity is currently under-represented by known structures in CATH. To compare the functional diversity of domain families of known structure in CATH and Gene3D, the functional diversity of each domain family was characterised by four criteria: (i) Annotation – this indicates whether there is one or more functional annotations for the family; (ii) Coverage – the percentage of members within a domain family with a functional annotation; (iii) Scope – the number of different functional annotations in the domain family; and (iv) Agreement – this indicates whether at least 80% of annotated family members have the same functional annotation. Average values for (ii) and (iii) are calculated for all families in Gene3D and CATH. In addition, the percentage of families classified as (i)

Annotated and (iv) in Agreement in Gene3D and CATH was also calculated. Figure 4.6 below illustrates these criteria.

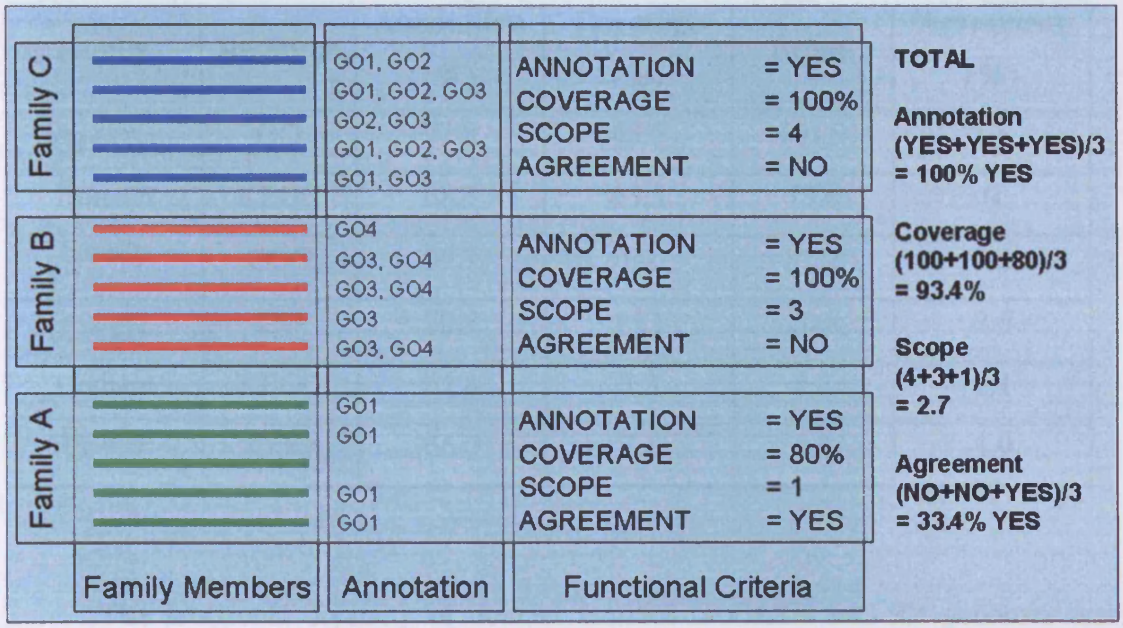


**Figure 4.6 Functional Characterisation of Domain Families.** *Characterisation of functional annotation in three domain families is illustrated. Note that scope is defined by the number of distinct, ordered, concatenated GO functional terms per family (i.e. in Family B "GO3, GO4", "GO3" and "GO4" gives a scope of 3).*

The functional characteristics of Gene3D domain families and CATH domain families are shown in table 4.3. Functional characteristics for the CATH classification were taken from the Dictionary of Homologous Superfamilies (DHS version 2.5.1, released January 2005, downloaded July 2005), which functionally annotates CATH (described earlier, see section 2.1.5.4). It can be seen that because Gene3D domain families have many additional sequence relatives from the genomes, there is a much clearer picture of the functional diversity than currently observed in CATH, with increased Annotation, Coverage and Scope.

**Table 4.3 Scope of Functional Annotation in Gene3D Families.** *Functional annotations in Gene3D and CATH domain families, assessed by Annotation, Coverage, Scope and Agreement.*

| Family | Resource | Annotation (%) | Coverage (%) | Scope | Agreement (%) |
|---|---|---|---|---|---|
| Gene3D Domain Family | GO | 97.8 | 45.5 | 191.9 | 1.5 |
| | KEGG | 68.3 | 63.5 | 73.6 | 0 |
| | COG | 69.7 | 1.2 | 82.2 | 0.2 |
| CATH Domain Family | GO | 70.9 | 15.1 | 20.8 | 2.6 |
| | KEGG | 62.8 | 7.3 | 6.7 | 2.1 |
| | COG | 66.0 | 7.4 | 6.9 | 1.9 |

The functional diversity of domain families in CATH and the genomes was compared by reference to GO functional annotations, since these have the highest coverage of the three resources (GO, KEGG, and COG). Functionally under-represented families were identified by the relative scope of the family between Gene3D and CATH. Relative scope was calculated by dividing the scope in Gene3D by the scope in CATH for each domain family. 380 domain families with a relative scope greater than 2 were considered functionally under-represented, representing 27,191 s35 subclusters.

Not all structurally under-represented domain families are also functionally under-represented. Of the 132 structurally under-represented domain families identified previously, 60 of these domain families are also functionally under-represented and thus can be selectively targeted as likely to be more informative for investigating the structural mechanisms for the evolution of function within a domain family.

The relationship between the sequence diversity and functional diversity of domain families in the genomes is shown in figure 4.7 below, revealing a general trend where sequence diverse domain families are also functionally diverse, supporting the observations of Reeves *et al.* and Ranea *et al.* (Reeves *et al.*, in preparation; Ranea *et al.*, submitted).

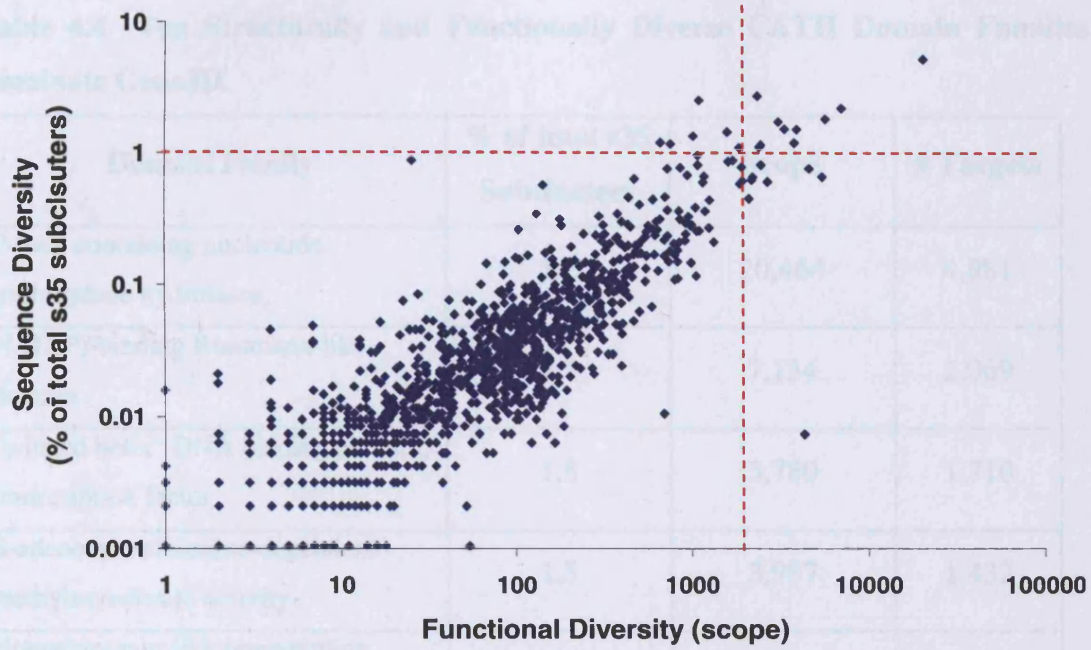


**Figure 4.7 Log-log plot showing Functional Diversity versus Sequence Diversity of CATH Domain Families in Gene3D.** *Functionally diverse families (greater than 2000 scope) and sequence diverse families (greater than 1% of total s35 subclusters) are indicated by red dotted lines.*

Ten domain families dominate the sequence diverse and functionally diverse CATH domain families in the genomes. Each of these families represents 1% or more of total s35 subclusters and has a functional scope of 2000 or more, shown in table 4.4 below. Collectively these ten domain families represent 20.1% of total s35 subclusters, and comprise 19,000 potential targets for structural genomics. Solving the additional structures from these families would enable greater understanding of the relationship between the structural variation and the evolution of functional diversity in these domain families. However, some of these families already dominate the PDB (for example the P-loop containing nucleotide triphosphate hydrolases). For these families we may want to use additional information in Gene3D to target those s35 subclusters which have particular GO functions for which we currently have no structural representative, or which have novel domain contexts for which we have no structural representative. Targeting novel domain contexts has been suggested previously by Vogel *et al.*, (Vogel *et al.*, 2004), target selection of novel domain contexts using Gene3D is currently ongoing.

**Table 4.4 Ten Structurally and Functionally Diverse CATH Domain Families Dominate Gene3D.**

| Domain Family | % of total s35 Subclusters | Scope | # Targets |
|---|---|---|---|
| P-loop containing nucleotide triphosphate hydrolases, | 5.2 | 20,464 | 4,881 |
| NAD(P)-binding Rossmann-like domain | 2.2 | 7,134 | 2,069 |
| "winged helix" DNA binding domain, transcription factor | 1.8 | 5,780 | 1,710 |
| S-adenosylmethionine-dependent methyltransferase activity | 1.5 | 3,987 | 1,432 |
| Homeodomain-like transcription regulation | 1.2 | 3,859 | 1,127 |
| Periplasmic binding protein-like II, transporter activity | 1.5 | 3,409 | 1,439 |
| Hydrolase activity, aromatic compound metabolism | 1.7 | 3,136 | 1,633 |
| Electron transport, Oxidoreductase activity, metabolism | 1.2 | 2,494 | 1,088 |
| Electron transport, Thioredoxin-like | 1.1 | 2,448 | 1,050 |
| Immunoglobulins | 2.7 | 2,388 | 2,571 |
| **TOTAL** | **20.1** | - | **19,000** |

## 4.4 Summary

Analysis using Gene3D can identify suitable targets for structural genomics initiatives. Of the 4,365 Pfam families identified in the genomes, that are structurally uncharacterised, selection of the largest 1,876 of these families permits a greater proportion of genome sequences to be structurally annotated. All these families contain more than 20 relatives, 447 contain more than 100 relatives, allowing multiple targets to be selected from these families to increase the probability that a substantial proportion of structures will be solved, given the high attrition rates.

The identification of 93,571 target subclusters from structurally characterised families in CATH would be required to accurately homology model all the relatives of domain families in the genomes from known fold groups. This is a prohibitively large number and underscores the need for rational prioritisation of structural genomics targets. Analysis using Gene3D has identified the most sequence diverse domain families, representing 20,763 target subclusters. Comparison of the distribution of sequence diversity between known structures classified in CATH and those predicted in the genomes has shown that many domain families are more sequence diverse than suggested by their current classification, with a 21-fold increase in the number of target clusters identified in Gene3D. 132 domain families have been identified that are significantly structurally under-represented, some of the most under-represented of these families perform diverse and biologically important functions, and greater insight into the structural mechanisms behind these functions may be gained from targeting the 5,966 homology modelling targets identified in these domain families. Analysis of the structural and functional diversity of domain families shows that ten domain families dominate the genomes - these families comprise 19,000 target subclusters.

In summary, Gene3D data allows sampling of all apparently diverse families to select targets for functions and domain contexts that have not yet been structurally characterised. By considering Kingdom distribution, the number of domain family relatives represented in each target subcluster, and the domain context of the domains represented in each target subcluster, further prioritisation of these targets can be accomplished.

# CHAPTER FIVE

# Phylogenetic Occurrence Profiles to Analyse the Function and Evolution of Domain Families

## 5.1 Introduction

Phylogenetic profiles have been exploited to detect functionally related proteins, and proteins that interact (see section 1.2.3.4 for a description and review of methods). Traditional phylogenetic profiles are based upon a presence/absence profile, whereby the presence of an orthologue in a genome is designated 1, and the absence 0. The resulting binary string represents the presence/absence profile of a protein across several genomes. These profiles can be compared and clustered into groups of profiles that are statistically significantly similar and may indicate that the proteins interact or are functionally related.

Bacterial genomes provide a good dataset for analysing protein domain and family evolution. The size of bacterial genomes has long been hypothesised to be under natural selection. The gene repertoire of bacterial genomes results from a balance between opposing mechanisms of gene increase, via horizontal transfer and gene duplication, and gene loss, via gene inactivation and deletion. The evolutionary mechanisms that maintain a small bacterial genome size, in order to maintain a competitive rapid rate of replication have been analysed by several researchers. Mira *et al.* (2001) describe a 'deletional bias' in bacterial genomes that acts to eliminate non-functional genes with decreased functional selection. The loss of genes that are no longer functionally selected for acts to maintain small bacterial genomes that lack non-functional sequences. Evolution by reduction, whereby the smallest genomes are derived from bacteria with larger genomes (Dobrindt and Hacker, 2001; Moran, 2002) has been described in many obligate intracellular bacterial pathogens. Massive gene loss significantly reducing genome size to an optimal genome size is an adaptive response to intracellular environmental selection pressures.

The gene repertoire of these genomes is far from static. As noted by Ochman *et al.* (2000) bacteria obtain a significant proportion of their genetic diversity through the acquisition of sequences from distantly related organisms. Horizontal transfer introduces substantial amounts of genetic material into bacterial genomes, combined with a high level of deletion of genetic material, bacterial gene repertoires are dynamic, maintaining a small efficient genome that can adapt to and exploit changing selective pressures. Bacterial genome size is not simply related to phenotype or lineage. Bacteria with a wide range of different phenotypes and lifestyles can have similar genome size; conversely, bacteria from narrow phylogenetic groups can have considerable diversity in genome size (Ochman *et al.*, 2000).

Gene duplication (Koonin *et al.*, 2002) and lineage-specific gene loss (Dobrindt and Hacker, 2001; Moran, 2002) are the primary mechanisms determining bacterial genome size. The influence of horizontal transfer in determining bacterial genome size is less apparent (Chothia *et al.*, 2003; Kurland *et al.*, 2003). Jordan *et al.* (2001), describe a positive correlation between the fraction of genes within lineage specific expansions (gene duplications occurring within specific prokaryotic lineages) and the total number of genes in a genome.

## 5.2 Objectives

The first section of this chapter describes using Gene3D to analyse bacterial evolution. In particular the identification of universal domains and the analysis of their genome frequency with regard to genome size. The following sections describe identification of genome size-dependent and universal domain families in bacteria using phylogenetic profiles, and the relationship between these domain families and bacterial genome size. Later sections of this chapter exploit domain family subclustering to generate more highly resolved phylogenetic profiles. A novel method of fine tuning these phylogenetic profiles to identify functionally related domain family subclusters is described, along with preliminary results describing some novel functional relationships identified using this method.

## 5.3 Results

There is a large complement of completely sequenced bacterial genomes in Gene3D. In collaboration with Juan Ranea, the distribution and occurrence of CATH homologous superfamilies in 100 bacterial genomes was investigated to explore the genetic and functional determinants involved in bacterial size distribution. We decided to study the distribution of CATH homologous superfamilies in bacteria, as there was a large dataset of bacterial genomes available. These genomes are less complex than those of eukaryotes (for example more accurate gene identification and annotation), which consisted of only 16 completed genomes existed at the time of this study. In bacterial genomes the absence of introns and long intergenic non-coding regions make open reading frame identification more accurate. Bacterial genome size will therefore be more likely to reflect the size of the bacterial proteome. Evolutionary selection on bacterial genome size, to promote small, reproductively efficient genomes, allows genome complexity and size to be estimated from the number of ORFs (Mira *et al.*, 2001), which is the measurement of genome size used in this chapter.

### 5.3.1 Analysis of the Genome Size Dependence of CATH Superfamilies

For each of the 940 CATH domains assigned to at least one gene in the bacterial genome dataset, an occurrence profile was derived (see figure 5.0). These occurrence profiles were used to explore the distribution of the CATH homologous superfamilies in the bacterial dataset.

| Domain Occurrence Profiles | | | | |
|---|---|---|---|---|
| Domain Superfamily: | Bacteria 1: | Bacteria 2: | Bacteria 3: | Bacteria 4: |
| A | 12 | 13 | 14 | 11 |
| B | 35 | 0 | 0 | 60 |
| C | 16 | 0 | 0 | 0 |

**Figure 5.0 Domain Occurrence Profiles.** *The number of open reading frames containing domain family relatives in each bacterial genome is converted into an occurrence profile for each CATH homologous superfamily.*

The occurrence of a CATH domain within each bacterial genome was calculated by counting the total number of different proteins in which the CATH domain had been assigned. Proteins that contained more than one CATH domain assignment of the same homologous superfamily were only counted once. This avoids redundancy and gives equal weight to domains occurring in different multi-domain contexts so more accurately reflecting a domains contribution to bacterial complexity. To explore the correlation between genome size and the number of domain relatives in a particular organism, correlation coefficients between genome size variation and CATH domain occurrence profiles were obtained by Spearman's rank correlation method (Kendall and Gibbons, Rank Correlation Methods ($5^{th}$ edition), 1990). 116 CATH domains with a Spearman's coefficient greater than or equal to 0.7 were considered significantly correlated with genome size and are termed size-dependent CATH homologous superfamilies (shown in figure 5.1). A Spearman's coefficient of 0.7 corresponds to a statistical probability of less than 0.0005 that the correlation occurs in the dataset by chance. These 116 size-dependent CATH homologous superfamilies represent only 12% of the total number of superfamilies occurring in the bacterial dataset, but account for 60% of domain assignments.

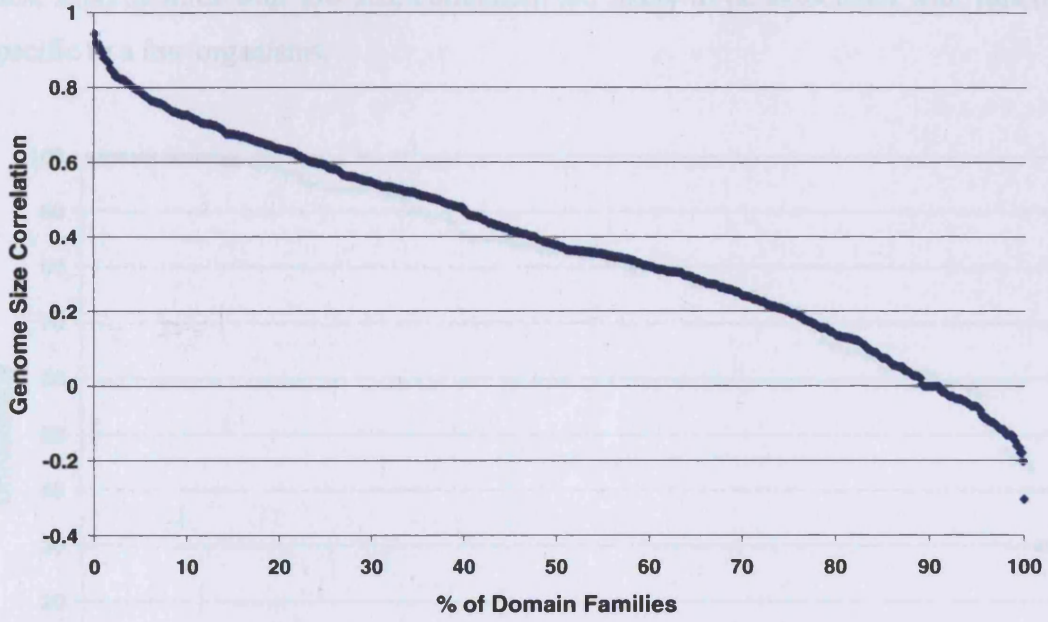


**Figure 5.1 Size-Dependent CATH Homologous Superfamilies.** *Spearman's Rank Correlation values for all CATH homologous superfamilies in the bacterial dataset. Correlation values greater then 0.7 indicate size-dependent superfamilies.*

### 5.3.2 Identification of Universal CATH Homologous Superfamilies

Within the set of size-dependent CATH homologous superfamilies identified in the bacterial dataset, a sub-set of universal CATH homologous superfamilies was identified. Size-dependent CATH homologous superfamilies with a wide representation in bacterial genomes (identified in at least 70% of genomes in the bacterial dataset) were considered universal to bacteria. A cut-off of 70% was considered an acceptable threshold since the sensitivity of the HMM method used to identify the CATH domains in the genomes is 76% using a structural dataset (see section 2.4.2.2). As can be seen from figure 5.2 below, 73% (85/116) of size-dependent CATH homologous superfamilies are universal to bacterial genomes. These size-dependent universal CATH homologous superfamilies represent only 9% of the total number of CATH homologous superfamilies, but account for 56% of domain assignments. This indicates that a few, highly recurring superfamilies are primarily responsible for genome complexity in bacteria, and suggests there are ancestral, homologous genetic mechanisms that narrowly specify genome complexity. Only 28% of non-size-dependent CATH homologous superfamilies are found to be universal, suggesting that

most superfamilies with low size correlation are likely to be associated with functions specific to a few organisms.

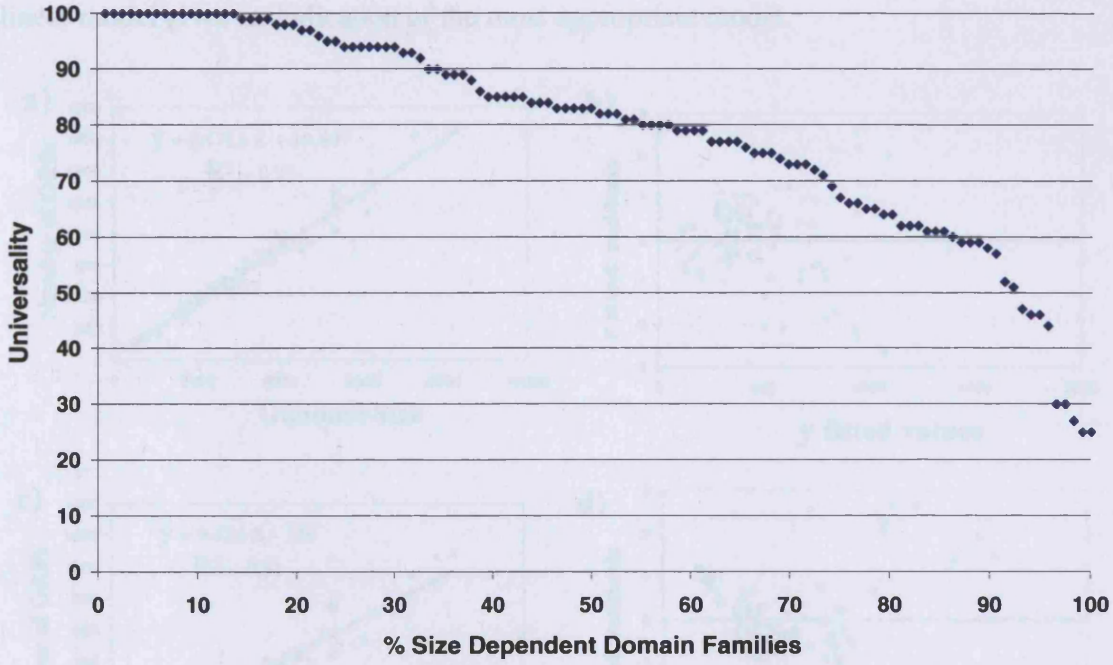


**Figure 5.2 Universal Size-Dependent Superfamilies.** *The universality value (percentage of total genomes containing domain family members) for all size dependent superfamilies in the bacterial dataset. Universal superfamilies are present in at least 70% of bacteria.*

### 5.3.3 Distribution of Size-Dependent Universal CATH Superfamilies

The distribution of 85 CATH homologous superfamilies that are both size-dependent and universal was analysed by plotting domain occurrence against genome size for each CATH domain in order to determine the type of relationship between domain occurrence and genome size. Linear regression model validation was performed by standardised residual analysis plotted against their respective y fitted values calculated from the linear regression model (Anscombe 1973; Atkinson 1985). Linear regression assumption was validated independently for all the universal and size-dependent superfamilies distributions. Three different types of distributions were identified: linear, power-law and logarithmic which characterised a total of 66 superfamilies. The type of distribution for the remaining 19 superfamilies was not clear. See appendix III for a complete list.

Figure 5.3 below shows the residual pattern analysis of the linear regression validation for the three main types of behaviour observed; linear, power-low and logarithmic. For each type of behaviour, the pattern of the error distribution around the linear model gives an indication of the most appropriate model.

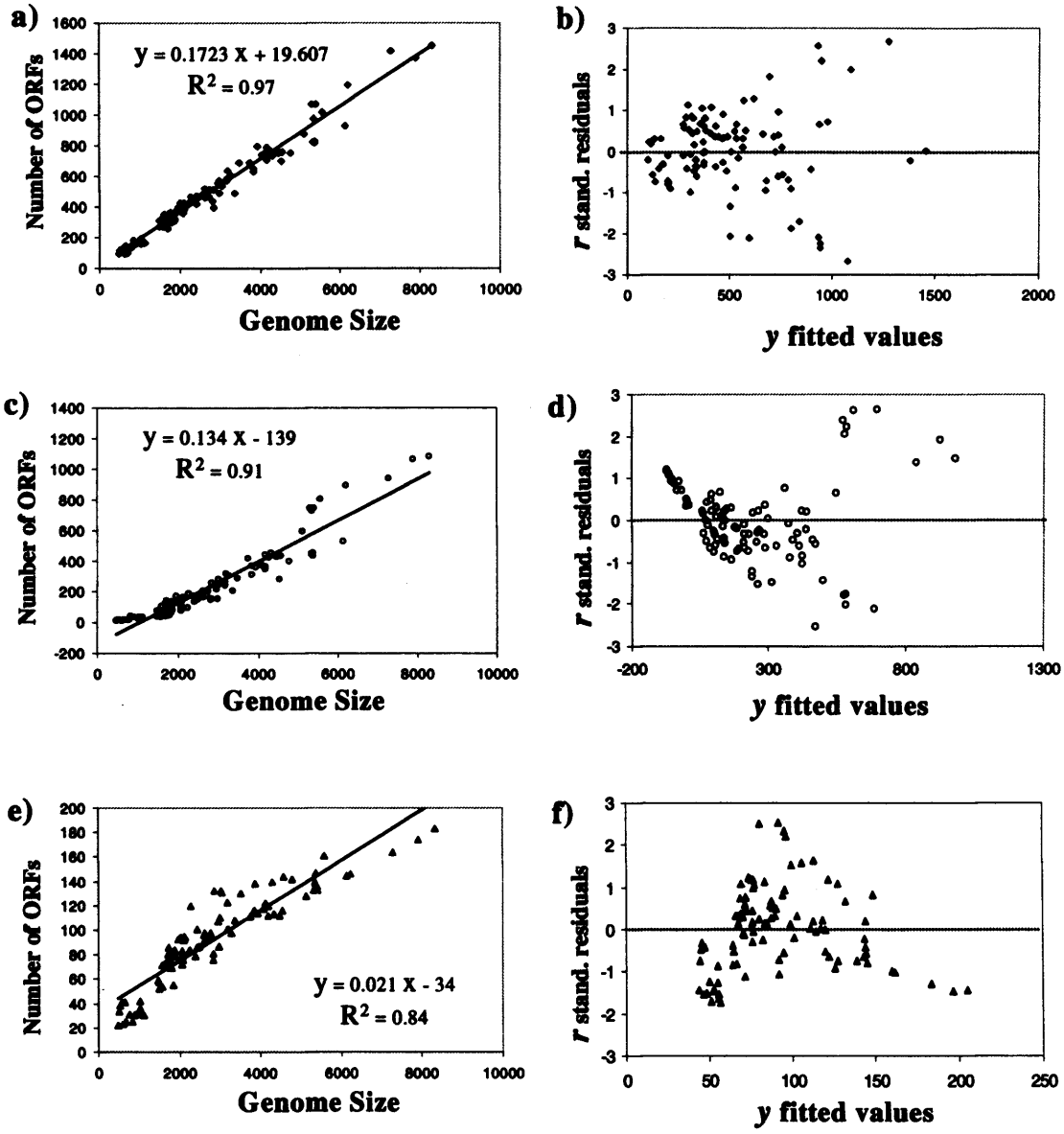


**Figure 5.3 Distribution of Size-Dependent Universal Superfamilies** *Plotting domain occurrences against genome size for each size-dependent universal superfamily reveals three types of distribution. Left hand plots show the best linear regression fitting for the three main types of domain superfamilies distributions: (a) linear, (c) power-law and (e) logarithmic. The equations for the regression lines and the $R^2$ values are indicated. The plots on the right show the standardised residual (r) (y-axis) calculated for each of the three linear regression models plotted against y^ fitted values (x-axis): (b) plot shows a residual distribution pattern confirming linearity, and (d) and (f) plots show curvatures characteristic of nonlinearity.*

### 5.3.4 Analysis of the Function of Size-Dependent Universal CATH Superfamilies

Functional analysis was undertaken for all superfamilies from each distribution set (linear, power law, logarithmic) using annotations from the COG database, as well as additional annotations from Pfam, SWISSPROT, CATH, SCOP and the literature where required. Functional classifications of each superfamily can be seen in appendix III.

The 38 linearly distributed, size-dependent universal superfamilies were found to be primarily involved with metabolism. 87% of domains and 82% of superfamilies in this category are involved in cellular metabolism. For example, the two most frequently occurring superfamilies in bacteria occur in this category: the nucleotide triphosphate hydrolase domain (CATH code 3.40.50.300) supplies reaction energy in both prokaryotes and eukaryotes; whilst the NADP-binding domain (CATH code 3.40.50.720) performs reducing or oxidising chemistry for a wide range of different reactions (Apic *et al.*, 2001; Hegyi *et al.*, 2002).

The 20 power law distributed, size-dependent universal superfamilies were found to be primarily involved in gene regulation mechanisms. 80% of domains and 60% of superfamilies in this category perform gene regulatory roles in bacteria. Major transcription factors occur in this category: winged-helix (CATH code 1.10.10.10); homeodomain-like (CATH code 1.10.10.60); and l-repressor DNA-binding domain (CATH code 1.10.160.10) (Babu and Teichmann, 2003). In addition to transcription factors, domains of the two-component signal transduction system (CheY response regulator domain (CATH code 3.40.50.2600), high-affinity periplasmic solute-binding protein (CATH code 3.40.190.10), and histidine kinase domain (CATH code 3.30.565.10)) are found in this category (Goudreau and Stock, 1998). Since regulatory domains show a high degree of modularity and usually combine with enzymatic or small molecule binding domains responsible for regulatory specificity (Apic *et al.*, 2001; Babu and Teichmann, 2003), it is perhaps not surprising that this category also contains some metabolic superfamilies.

The 8 logarithmically distributed, size dependent universal superfamilies do not show any common functional tendency. These superfamilies are involved in diverse

functions such as metabolism, RNA binding and DNA repair, which makes it difficult to define a single functional term for this category. However, since these superfamilies represent only 10% of all size dependent universal domains and 9% of all size dependent universal superfamilies they represent only a small proportion of superfamilies and are not considered for further analysis.

### 5.3.5 Identifying the Bacterial Genome Size Determinants of Size Dependent Universal Superfamilies

When considering the influence of size dependent universal superfamilies upon bacterial genome size, the balance between metabolic and regulatory functions and their respective costs can be considered. As hypothesised by Bird (1995), increasing complexity is limited by the increase in logistical problems of distinguishing signal from noise. In terms of bacterial complexity, the benefits of environmental exploitation and adaptation gained from an increase in genome size come with combinatorial increases in protein network complexity to fully integrate and apply the functions of additional genes. As each additional gene is added to a genome, the combination of possible gene interactions and protein interactions increases by an increasing amount. Thus the benefits derived from expansion of the metabolic repertoire are accompanied by an increase in the regulatory repertoire necessary to control interactions between these proteins and so distinguish signal from noise.

#### 5.3.5.1 Economies of Scale

The relationship between a linear increase in one factor (i.e. metabolic repertoire) requiring a non-linear increase in another (i.e. regulatory repertoire), is analogous to the factory optimisation model exploited in economic analyses (Mankiw, Principles of Microeconomics (2nd edition), 1998). The factory optimisation model is used to describe microeconomics – a single factory which produces a single product which is always in demand. In the factory optimisation model illustrated in figure 5.4, a linear increase in unit production produces a non-linear increase in production overheads. At low levels of unit production there is little restraint on production unit increases as the required increases in production overheads are low. However, at high

levels of unit production there is a large restraint on production unit increases as the required increases in production overhead cost more than the benefit of increased unit production. A major factor influencing optimum size in productive systems is the effect where any linear increase in production complexity is usually associated with a larger increase in the associated logistic cost (Frizelle, The Management of Complexity in Manufacturing, 1998; Orr, 2000).

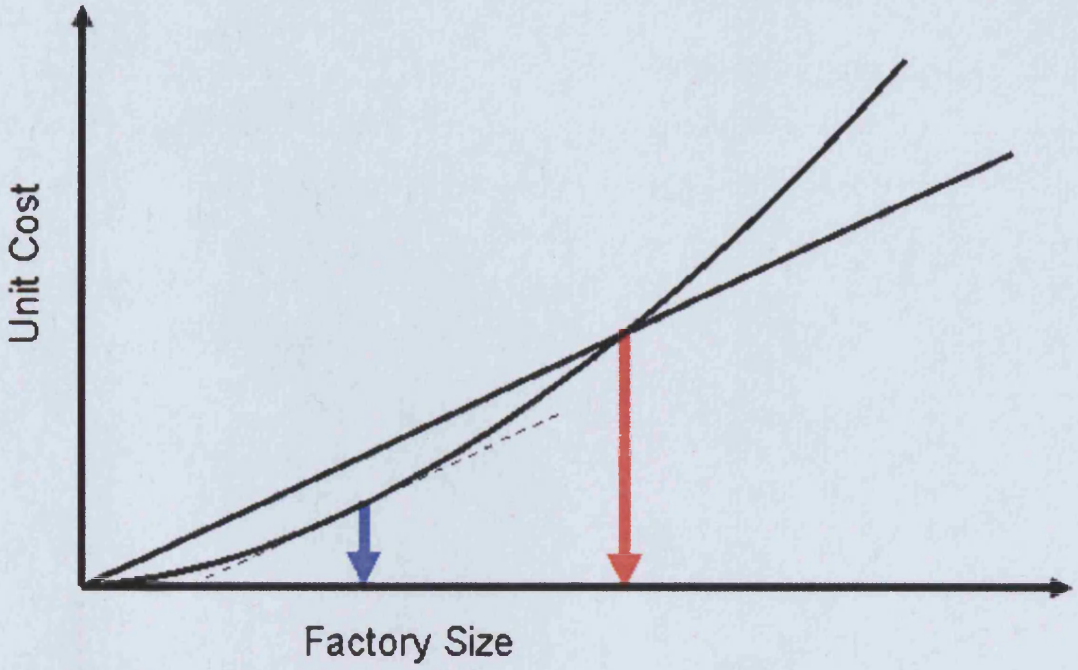


**Figure 5.4 Economies of Scale: Optimum Factory Size** *Factory production revenue (straight line) and factory production overhead costs (curve) values show maximum economic factory size (red arrow) where any further increase in production revenue results in a larger increase in production overhead costs, making an increase in factory size unprofitable; and optimal factory size (blue arrow) where production revenue is offset by an equivalent production overhead cost, maximising profit.*

5.3.5.2 Predicting Optimal Bacterial Genome Size

This model can be applied to an analysis of bacterial genome size, shown in figure 5.5. An analogy between unit production and metabolic capacity can be made since an increase in metabolic capacity provides bacteria with new ways to exploit the environment and hence increase survivability. The regulatory processes required to

integrate and control metabolic systems is analogous to the cost of production overheads, since an increase in the number of regulatory genes can be considered a cost because bacteria have to keep their genome size to a minimum in order to replicate most efficiently, which is important for their survival. The cellular cost of regulatory systems can be seen in bacteria with stable substrate sources (for example endosymbiotic bacteria) where the strategy of constant expression of metabolic enzymes and loss of regulatory genes is observed. In contrast, in free living bacteria, efficient exploitation and response to diverse environmental pressures provides a survivability advantage that offsets the cost of maintaining regulatory systems.

Fitting the distributions of metabolic (linearly distributed size-dependent universal superfamilies) and regulatory (power law distributed size-dependent universal superfamilies) superfamilies to regression lines shows two interesting effects of these superfamilies on bacterial genome size.

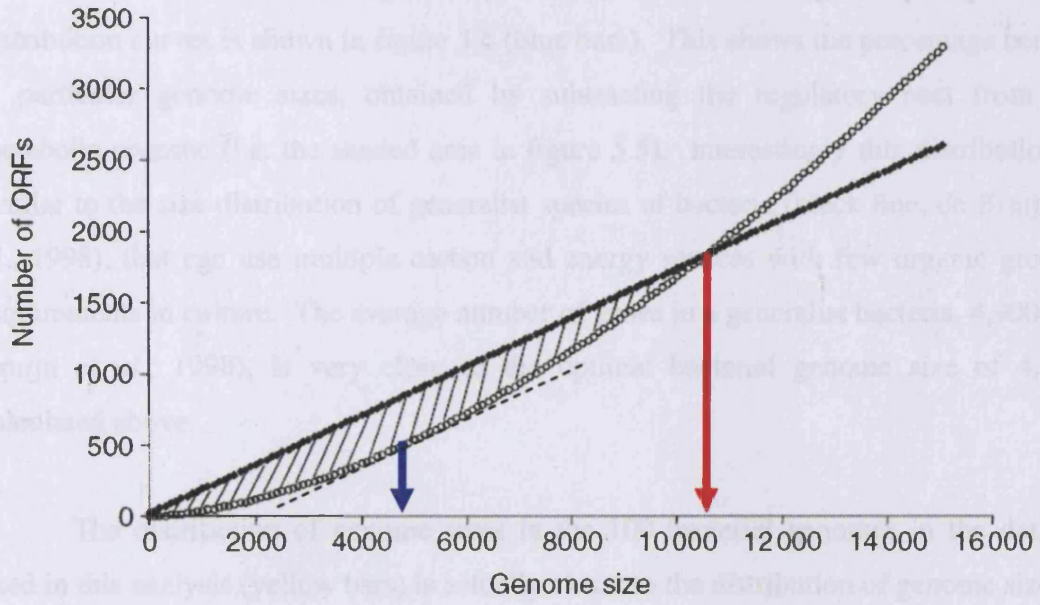


**Figure 5.5 Economies of Scale: Optimum Bacterial Genome Size.** *Plotting the distribution curves of metabolic size-dependent universal superfamilies and regulatory size-dependent universal superfamilies indicates maximum bacterial genome size (red arrow) and optimal bacterial genome size (blue arrow).*

The two functions cross at 10,500 open reading frames. This suggests that above a genome size of 10,500 ORFs the cost of additional regulation exceeds the benefit of increased metabolic complexity. The optimal bacterial genome size occurs when the incremental costs of metabolic increase and regulatory increase are equal, this level of metabolic complexity is of maximal efficiency, since it is achieved at minimal relative regulatory cost, i.e. where the regulatory superfamilies distribution gradient is equal to the metabolic superfamilies distribution gradient. This optimum bacterial genome size occurs at 4,805 open reading frames. Deviation away from this optimum incurs a cost of reduced reproductive efficiency. Above optimal genome size an increase in metabolic complexity demands a comparatively higher increase in regulatory complexity. For example, as can be seen from the gradient of the regulatory superfamilies distribution curve, the regulatory increment required to add one gene when the genome size is 8000 is almost triple that required when the genome size is 2000.

The size distribution given by the metabolic and regulatory superfamily distribution curves is shown in figure 5.6 (blue bars). This shows the percentage benefit at particular genome sizes, obtained by subtracting the regulatory cost from the metabolic revenue (i.e. the shaded area in figure 5.5). Interestingly this distribution is similar to the size distribution of generalist species of bacteria (black line, de Bruijn *et al.*, 1998), that can use multiple carbon and energy sources with few organic growth requirements in culture. The average number of genes in a generalist bacteria, 4,900 (de Bruijn *et al.*, 1998), is very close to the optimal bacterial genome size of 4,805 calculated above.

The distribution of genome sizes in the 100 bacterial genomes in the dataset used in this analysis (yellow bars) is actually closer to the distribution of genome size of specialist bacteria (intracellular bacteria incapable of reproducing by themselves; dotted line, de Bruijn *et al.*, 1998) than generalist bacteria (free living bacteria). The fact that the calculated optimal genome size derived from metabolic and regulatory superfamily distribution curves is in good agreement with the average genome size of generalist bacteria, indicates that the size dependent universal superfamilies represent universal molecular technology shared by all prokaryotes to perform their metabolic and regulatory processes, and that all bacteria have used similar molecular technology to optimise their reproductive efficiency. This efficiency has been achieved by maintaining

a metabolic complexity and associated regulatory complexity that balance the capacity for environmental exploitation with reproductive efficiency.

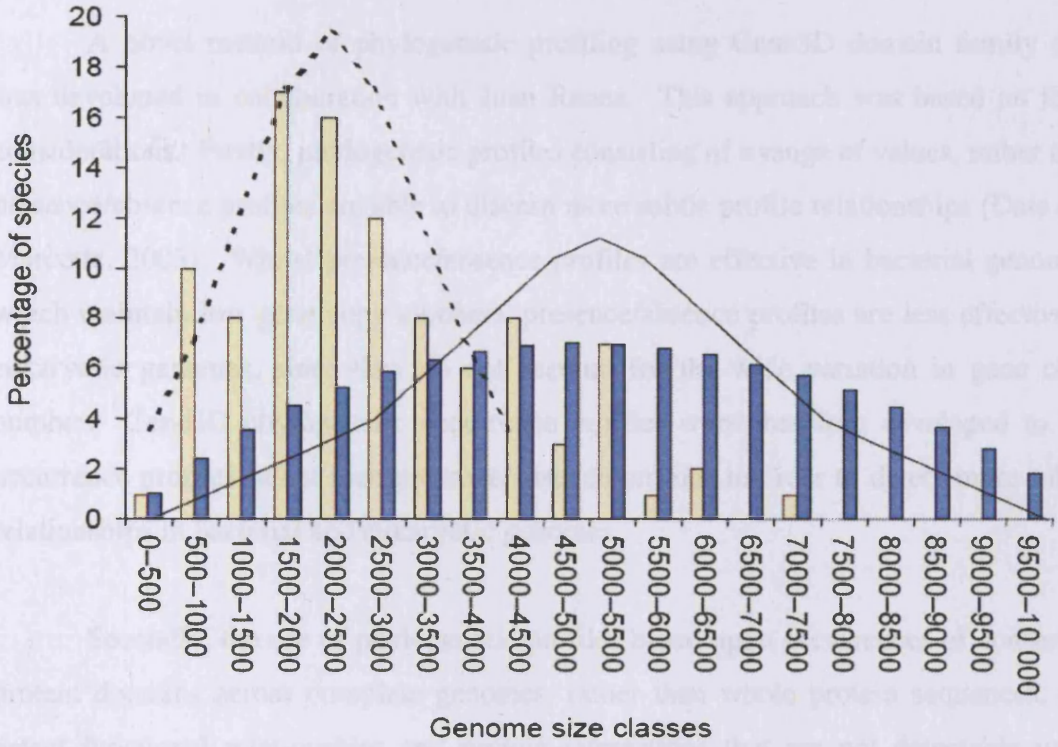


**Figure 5.6 Distribution of Bacterial Genome Size.** *The size distribution curve for bacterial genomes predicted from metabolic/regulatory size-dependent universal superfamilies (figure 5.5 hashed area) shown by blue bars is similar in range and shape to the size distribution of generalist bacteria (black line). The size distribution of the 100 bacteria in the dataset (white bars) is similar to the size distribution of specialist bacteria (dotted line).*

## 5.4 Using Gene3D Phylogenetic Occurrence Profiles for Predicting Protein Functional Relationships

A novel method of phylogenetic profiling using Gene3D domain family data was developed in collaboration with Juan Ranea. This approach was based on three considerations. Firstly, phylogenetic profiles consisting of a range of values, rather than presence/absence profiles are able to discern more subtle profile relationships (Date and Marcotte, 2003). Whilst presence/absence profiles are effective in bacterial genomes, which maintain low gene copy numbers, presence/absence profiles are less effective in eukaryotic genomes, since they do not account for the wide variation in gene copy number. Gene3D phylogenetic occurrence profiles were therefore developed to use occurrence profiles rather than presence/absence profiles in order to detect more subtle relationships in bacterial and eukaryotic genomes.

Secondly, the use of phylogenetic profiles based upon occurrences of conserved protein domains across complete genomes, rather than whole protein sequences, can detect functional relationships and protein interactions that are not detectable using phylogenetic profiles of whole proteins (Pagel *et al.*, 2004). Gene3D phylogenetic profiles use the occurrence of CATH homologous superfamilies across complete genomes. At present, only CATH domain assignments are used to generate phylogenetic occurrence profiles, an obvious future expansion of this method would be to incorporate Pfam, Newfam and other domain assignments (via InterPro).

Thirdly, and perhaps most importantly, Gene3D phylogenetic occurrence profiles are derived not only from the occurrence of CATH homologous superfamilies across complete genomes, but since CATH homologous superfamily assignments have been clustered into domain family subclusters at several different levels of sequence identity (for example s30, s35, s40, s50, s60, s70, s80, s90, s95 and s100 subclusters, described previously, see section 4.3.1.2), phylogenetic profiles are also constructed from each domain family subcluster level. This unique approach allows individual domain family subcluster profiles to be compared and clustered into statistically significant groups of profiles, thus identifying functional relationships and protein interactions between specific domain family subclusters. The building of Gene3D phylogenetic profiles is illustrated in figure 5.7 below.

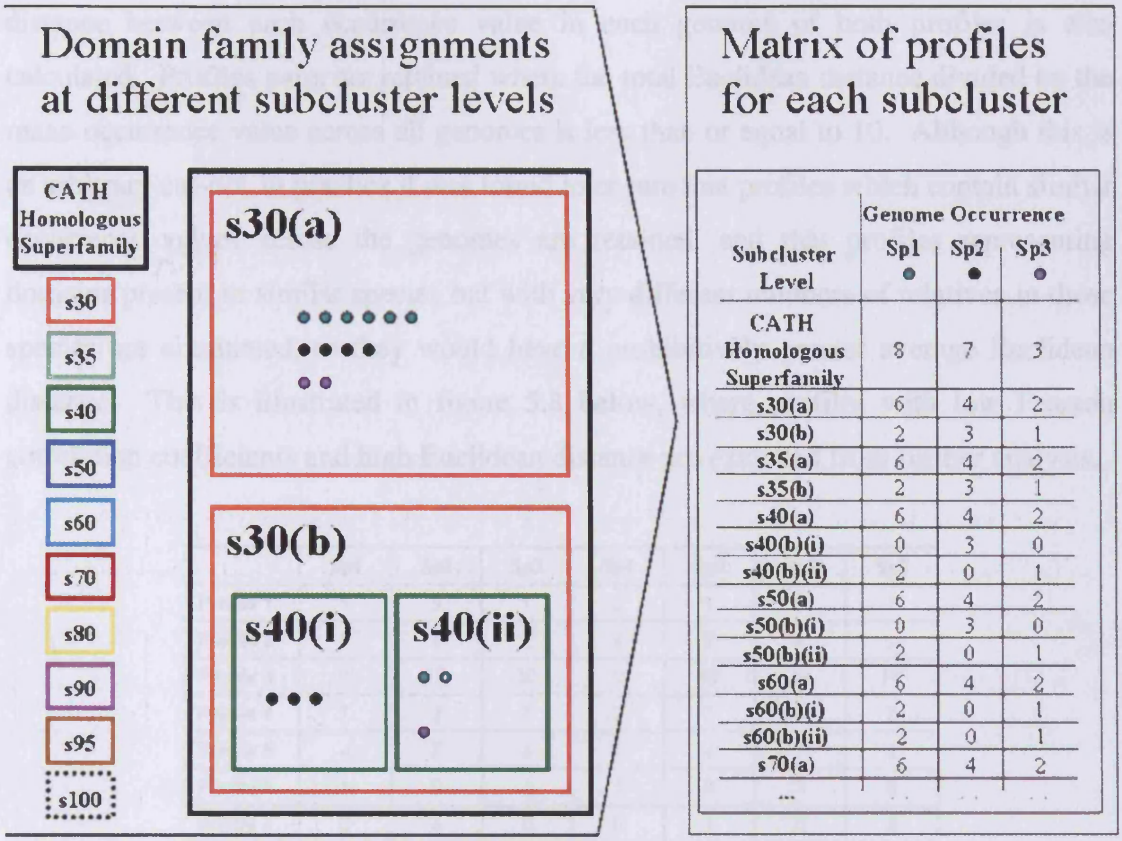


**Figure 5.7 Building Gene3D Phylogenetic Profiles.** *An occurrence profile is derived from the number of relatives identified in each genome from each CATH domain family and from each CATH domain family subcluster level within the domain family. Thus multiple profiles are generated from a single domain family that represent the occurrence of domain family relatives and the occurrence of domain family subcluster relatives across complete genomes in Gene3D.*

### 5.4.1 Pair Comparison of Profiles

Once phylogenetic profiles have been built for all CATH domain families and domain family subclusters in Gene3D, the similarity between each profile and all other profiles is calculated by an all versus all pairwise profile comparison, which uses two similarity criteria.

Firstly, for each profile pair, the Pearson correlation coefficient (Weisstein, 1999) is calculated, pairs with a Pearson correlation coefficient of 0.8 or higher are

retained since this indicates significant similarity between these profiles. The Euclidean distance between each occurrence value in each genome of both profiles is also calculated. Profiles pairs are retained where the total Euclidean distance divided by the mean occurrence value across all genomes is less than or equal to 10. Although this is an arbitrary cut-off, in practice it was found to ensure that profiles which contain similar occurrence values across the genomes are retained, and that profiles representing domains present in similar species but with very different numbers of relatives in those species are eliminated, as they would have a prohibitively greater average Euclidean distance. This is illustrated in figure 5.8 below, where profiles with low Pearson correlation coefficients and high Euclidean distance are excluded from further analysis.

| | Sp1 | Sp2 | Sp3 | Sp4 | Sp5 | Sp6 | Sp7 |
|---|---|---|---|---|---|---|---|
| Profile 1 | 1 | 3 | 1 | 3 | 1 | 3 | 1 |
| Profile 2 | 2 | 4 | 2 | 4 | 2 | 4 | 2 |
| Profile 3 | 10 | 13 | 10 | 13 | 10 | 13 | 10 |
| Profile 4 | 7 | 2 | 7 | 2 | 7 | 2 | 7 |
| Profile 5 | 4 | 7 | 4 | 7 | 4 | 7 | 4 |
| Profile 6 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| Profile 7 | 0 | 4 | 0 | 0 | 1 | 0 | 3 |

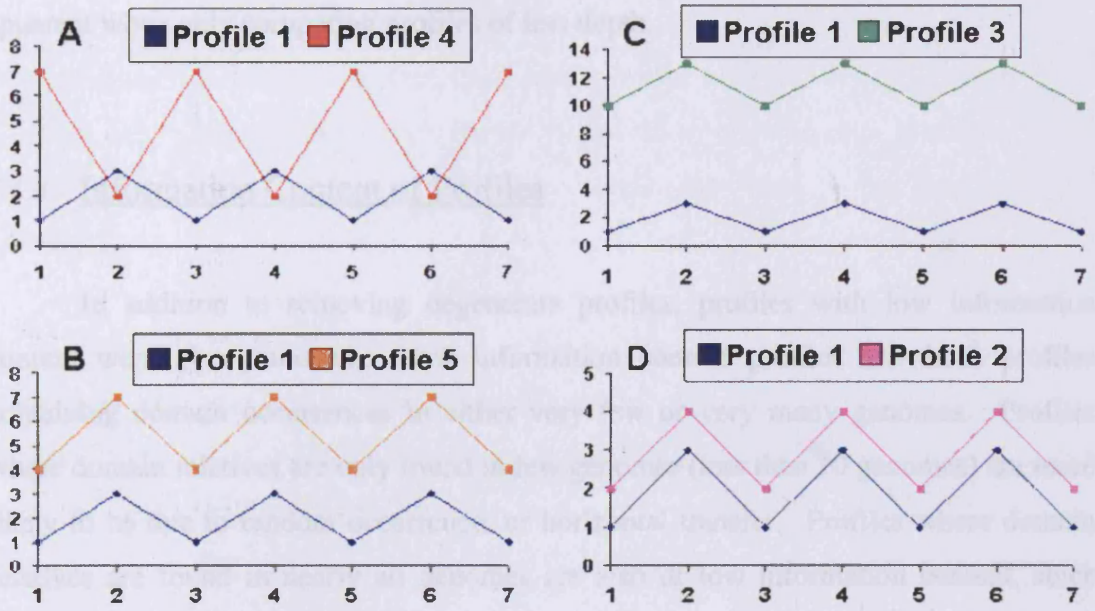


**Figure 5.8 Profile Pair Comparison.** *An all versus all pairwise profile comparison selects profile pairs that have a Pearson correlation coefficient of 0.8 or higher (for example profile pair 1 and 4 (A) is rejected ($P_1$ <0.8) whilst profile pair 1 and 5 (B) is retained ($P_2$ > 0.8)), and a Euclidean distance/mean occurrence value less than or equal to 10 (for example profile pair 1 and 3 (C) is now rejected ($E_1$ > 10) whilst profile pair 1 and 2 (D) is retained ($E_2$ < 10)).*

### 5.4.2 Degenerate Domain Family Subcluster Profiles

Gene3D profiles pairs that are retained (according to the selection criteria described above, see section 5.4.1) are further filtered to remove profile pairs that are degenerate, that is where a profile pair between profiles from a deeper subcluster level has already been retained. This reduces the number of profiles without reducing the correlation information they represent, and thus enables faster profile clustering (discussed below). For example, in a profile pair consisting of profile A (representing a domain family 1.10.10.10 s40 subcluster) and profile B (representing a domain family 3.40.50.300 homologous superfamily); this profile pair is degenerate and can be removed if another profile pair has been retained containing profiles from 'deeper' subcluster levels. For example a pair consisting of the same profile A but a deeper profile subcluster than profile B (for example representing a domain family 3.40.50.300 s35 subcluster profile, instead of the homologous superfamily). This filtering does not remove any informative pairwise relationships, but permits 'fine tuning' of any identified correlated profile pairs to the deepest domain family subcluster level, and so may reveal correlations between different domain family members that may not be apparent when only comparing profiles of less depth.

### 5.4.3 Information Content of Profiles

In addition to removing degenerate profiles, profiles with low information content were also removed. Low information content profiles are those profiles containing domain occurrences in either very few or very many genomes. Profiles where domain relatives are only found in few genomes (less than 10 genomes) are more likely to be due to random occurrence, or horizontal transfer. Profiles where domain relatives are found in nearly all genomes are also of low information content, since these profiles do not show any distinctive phylogenetic pattern. Removal of these low information content profiles reduces the number of false-positive profile relationships (Pagel *et al.*, 2004).

### 5.4.4 Comparison of Gene3D Profiles to Randomised Null Models

The functional relevance of related profiles derived from Gene3D data could be obscured by correlations due to non-functional factors in our dataset such as correlations between profiles resulting from variations in the size of genomes in Gene3D or variations in the size of domain families in Gene3D. To account for these factors, Gene3D data was randomised in two different ways to generate two different null model datasets: genome shuffling and profile shuffling.

The genome shuffling null model dataset is used to estimate the effect that genome size has on profile correlations. This dataset is generated by randomly shuffling domain assignments between genomes, each genome receives the same number of domain assignments it had before, but because these assignments are made randomly, any correlations due to genome size are eliminated.

The profile shuffling null model dataset is used to estimate the effect of domain family size on profile correlations. This dataset was generated by shuffling each domain family s100 subcluster profile, in effect assigning domain occurrences to random domain families. All other lower level profiles were then regenerated from this shuffled data. The resulting null model dataset effectively contains genomes with a generic set of domain assignments where any correlations due to domain family size are eliminated.

Pairwise Pearson correlation coefficients for each null model dataset were calculated, and the resulting distribution of Pearson correlation coefficients compared to Gene3D profile data. As can be seen in figure 5.9 below, neither of the null model datasets produced pairwise comparisons with significant Pearson correlation coefficients of 0.8 or higher, indicating that Gene3D profile pairs with Pearson correlation coefficients above 0.8 are likely to be due to the correlated evolution of functionally related or interacting protein domains, rather than effects of genome size or domain family size.

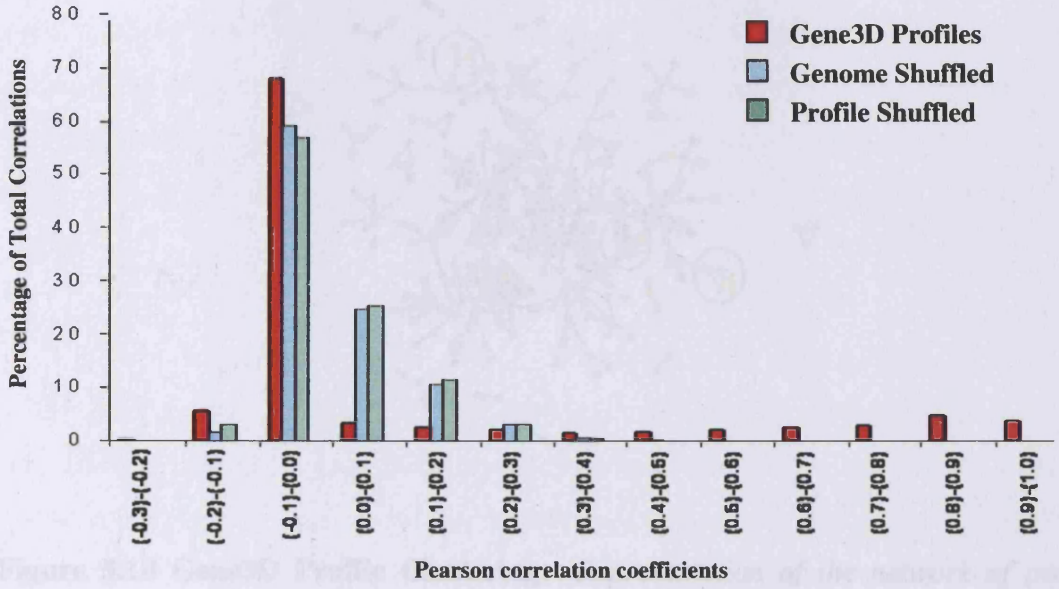


**Figure 5.9 Comparison of Gene3D Profile Correlations to Randomised Null Model Datasets.** *The distribution of Pearson correlation coefficients in all versus all pairwise profile comparison for each of three datasets: Gene3D profiles (red), genome shuffled null model profiles (blue) and profile shuffled null model profiles (green) is shown.*

### 5.4.5 Gene3D Phylogenetic Occurrence Profile Clustering

Gene3D profiles are clustered into single linkage and multilinkage clusters on the basis of their Pearson correlation coefficients, using TCluster (described previously, see section 2.3.3). Multilinkage clustering produces tight clusters where all the member profiles in a cluster have a significant relationship to one another. This produces quite restrictive clusters, often containing less than ten member profiles (see figure 5.10 below). Single linkage clustering produces much less restrictive clusters, since a member profile within a cluster only requires one significant relationship to an existing cluster member to be included in the cluster. As can be seen in figure 5.10 below, this can result in a giant cluster containing the vast majority of profiles. This giant cluster effect is typical of small world networks which have been observed in other biological data including protein domain co-occurrence, protein-protein interactions, and metabolic pathways (Wuchty and Almaas, 2005).

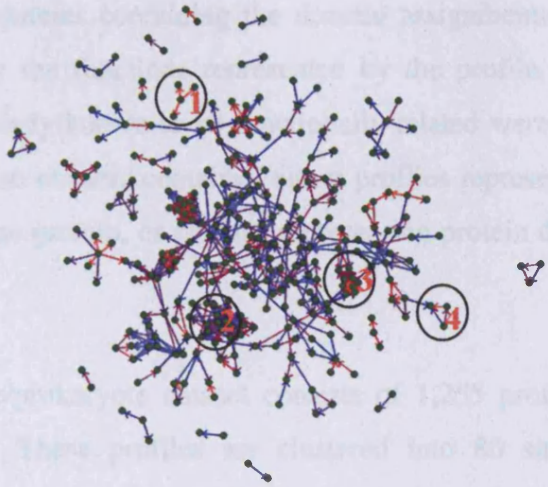


**Figure 5.10 Gene3D Profile Clustering.** *Representation of the network of profile clusters seen using single linkage. The figure shows clustering of eukaryotic profiles from Gene3D produced using the BioLayout program (Goldovsky et al., 2005). Labelled clusters indicate profiles representing proteins involved in 1: Actin and VCP-like ATPases, 2: Chaperones and Cytoskeleton, 3: DNA Replication and Repair, and 4: DNA Topoisomerase and Elongation Factor G.*

### 5.4.6 Functional Clusters revealed by Gene3D Phylogenetic Occurrence Profile Clustering

The 1277 CATH homologous superfamilies assigned in Gene3D produced 2.1 million individual domain family and domain family subcluster profiles, providing 2.2 billion possible pairwise profile comparisons. These profiles were analysed in three analysis groups: eukaryotic profiles, prokaryotic profiles and eukaryotic/prokaryotic profiles. After the filtering process, the number of profiles with significant pairwise relationships is significantly reduced. For example, the eukaryote/prokaryote analysis group contains 1,255 profiles found in 4,029 profile pairs.

#### 5.4.6.1 Profile Clusters Representing known Functional Groups

The functional significance of clusters was analysed using GO functional information from Gene3D. For each profile in a cluster, the GO functional terms

associated with the proteins containing the domain assignments that built the profile were used to identify the functions represented by the profile. Profile clusters that represent proteins already known to be functionally related were identified in all three analysis groups. These clusters contained either profiles representing different protein domains from the same protein, or profiles representing protein domains from different proteins.

The eukaryote/prokaryote dataset consists of 1,255 profiles that are found in 4,029 profile pairs. These profiles are clustered into 80 single linkage and 214 multilinkage clusters. An example identified in eukaryotic/prokaryote multilinkage clustering is the urea amidohydrolase multilinkage cluster, which contains four profiles, all of which represent domains occurring in urea amidohydrolase. Urea amidohydrolase catalyses the hydrolysis of urea to ammonia and carbamate. The enzyme consists of three protein subunits: alpha, beta and gamma (Mobley *et al.*, 1995). Profiles representing both the urease alpha domain (CATH code 2.30.40.10) and the metal-dependent hydrolase domain (CATH code 3.20.20.140) in the urease alpha protein cluster with profiles representing the urease beta domain (CATH code 2.10.150.10) in the urease beta protein and the urease gamma domain (CATH code 3.30.280.10) in the urease gamma protein. Profiles in this cluster represent domain family relatives in 36 genomes, of which 2 are eukaryotic (*A. thaliana* and *S. pombe*) and 34 are prokaryotic. Interestingly, whilst the domain profiles from urease beta and gamma subunits are homologous superfamily level profiles, the profiles representing the two domains in the urease alpha subunit are both domain family s30 subcluster profiles, implying that these specific domain family s30 subclusters are functionally and evolutionarily more closely linked, as might be expected from domains occurring in the same protein.

Initial analysis of the eukaryotic dataset, single linkage clustering produced a giant cluster, as can be seen from figure 5.10 above. In addition, several interesting clusters were identified, containing profiles representing Actin/VCP-like ATPases; Chaperones/Cytoskeleton; DNA Replication/Repair; and DNA Topoisomerase /Elongation Factor G. These clusters can be seen in figure 5.10 above. The DNA topoisomerase/elongation factor G cluster contains three profiles, shown in figure 5.11 below.

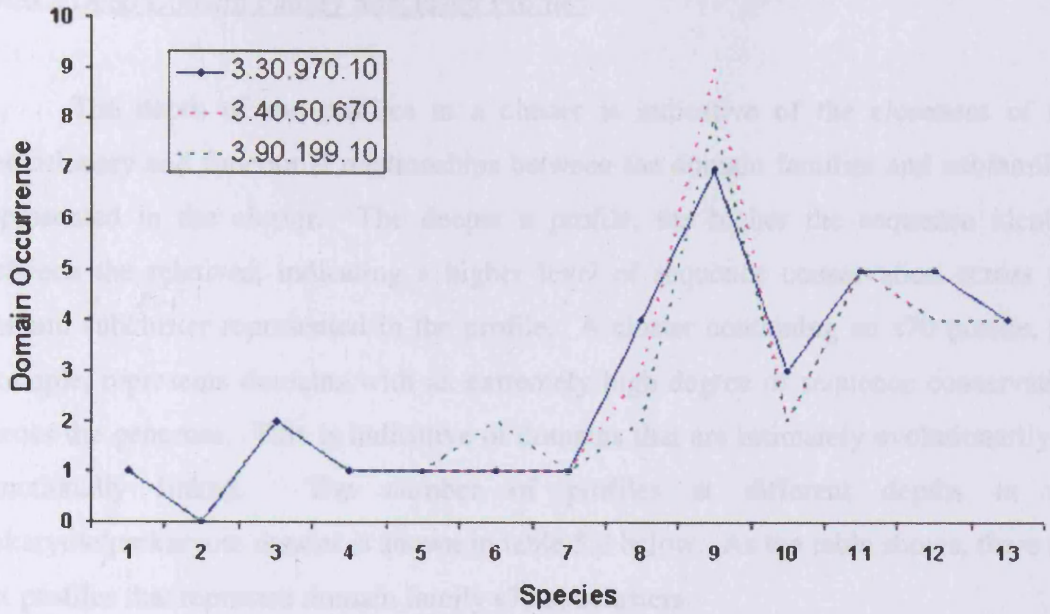


**Figure 5.11 Eukaryotic Cluster Domain Occurrence Profiles.** *Domain occurrence in each species represented in the eukaryotic dataset for the three profiles in the DNA Topoisomerase/Elongation Factor G cluster.*

Two of these profiles represent domains found in the same protein, elongation factor G (CATH codes 3.40.50.670 and 3.90.199.10), whilst the remaining profile represents a domain found in DNA topoisomerase (CATH code 3.30.970.10). The domain occurrences for each profile in the cluster are shown in figure 5.11 below. Both proteins represented by profiles in this cluster are involved in protein synthesis, whilst elongation factor G catalyses the translation reaction in the ribosome, DNA topoisomerase unpackages/repackages the DNA. Experimental evidence suggests that there may be a functional association between DNA topoisomerase and protein translation regulation. Rapisarda *et al.*, (2004), describe the repression of hypoxia-inducible factor 1 (HIF-1) protein accumulation by the topoisomerase poison topotecan (TPT). The authors show that TPT inhibits HIF-1 translation, and that topoisomerase is required for this inhibition. The authors conclude that a novel pathway connects topoisomerase-dependent signalling events and the regulation of HIF-1 protein expression and function. The identification of a cluster containing domains from topoisomerase and elongation factor G provides potential insight into the mechanisms in this novel pathway.

5.4.6.2 Deep Domain Family Subcluster Profiles

The depth of the profiles in a cluster is indicative of the closeness of the evolutionary and functional relationships between the domain families and subfamilies represented in the cluster. The deeper a profile, the higher the sequence identity between the relatives, indicating a higher level of sequence conservation across the domain subcluster represented in the profile. A cluster containing an s70 profile, for example, represents domains with an extremely high degree of sequence conservation across the genomes. This is indicative of domains that are intimately evolutionarily or functionally linked. The number of profiles at different depths in the eukaryote/prokaryote dataset is shown in table 5.0 below. As the table shows, there are six profiles that represent domain family s70 subclusters.

**Table 5.0 Depth of Profiles in Eukaryote/Prokaryote Dataset.** *The number of profiles of each depth (subcluster level) is indicated. Note that there are no profiles that represent domain family subclusters levels deeper than s70.*

| Profile Depth | Frequency |
|---|---|
| Homologous Superfamily | 39 |
| s30 | 193 |
| s35 | 109 |
| s40 | 78 |
| s50 | 45 |
| s60 | 14 |
| s70 | 6 |

This section describes one of the clusters containing a domain family s70 subcluster profile in the eukaryote/prokaryote dataset, which can be used to predict probable functional relationships of profiles representing domains in proteins that have no known function. The cluster contained two profiles. One profile represented a domain family s70 subcluster of the CATH homologous superfamily 2.20.29.10, functionally annotated as 'translocation elongation factor G' (a G protein factor that catalyses the translocation of peptidyl-tRNA from the A site to the P site of the ribosome during protein synthesis). The other profile in the cluster represented a domain family s30 subcluster of CATH homologous superfamily 3.40.50.610, functionally annotated as 'conserved hypothetical protein'. Both these profiles have identical species distributions across 32 genomes, and represent 37 domain assignments

to 37 different proteins across the 32 genomes, and 34 domain assignments to 34 different proteins across the 32 genomes respectively. These domain assignments occurred in different proteins in each genome, indicating that within each of the 32 genomes, these proteins were functionally associated and in addition, that the 34 conserved hypothetical proteins were likely to be specifically functionally associated with a highly conserved 2.20.29.10 domain family s70 subcluster found only in elongation factor G. Although the proteins containing domains represented in the domain family s30 subcluster have no known function, and are simply annotated as 'conserved hypothetical proteins', Gene3D phylogenetic profile clustering provides compelling evidence that predicts these proteins are intimately associated with elongation factor G, an essential protein involved in the elongation process during protein synthesis.

### 5.4.7 User Defined Query Profiles

Gene3D profile clusters can also be searched to identify proteins that have a functional relationship to user-defined data. In a test example, Gene3D profiles were searched to find proteins that may have a significant functional relationship to *ras* (a well studied protein with an established role in eukaryotic signalling cascades). Query phylogenetic profiles were generated using *ras* occurrences across eight eukaryotic genomes. In these profiles the user defined the subset of, in this case, eight species which made up the occurrence profile, and manually assigned *ras* occurrences to these genomes. However, the user could have chosen any subset of species from which to generate the query profile, and any method of defining protein or domain occurrence in their chosen genomes. Gene3D profiles were searched to identify significantly similar Gene3D profiles to the query profile, by calculating Pearson correlation coefficients and Euclidian distances between the *ras* query profiles and Gene3D profiles. Single linkage clustering of eukaryotic profiles and the user-defined *ras* query profiles produced a *ras* cluster containing the *ras* query profile and 17 additional profiles, these are shown in table 5.1 below.

From table 5.1 it can be seen that three profiles in the *ras* cluster represent proteins directly involved in signal transduction and the activation of transcription. Six profiles represent proteins involved in transcriptional regulation, including developmental transcription regulators with Homeobox domains and Polycomb

domains. Proteins with these domains are already known to be involved in embryonic development and regulation of cell cycle control acting through *ras*–mediated signalling pathways (Maclean *et al.*, 2005; Jacobs *et al.*, 1999).

The presence of profiles representing proteasome 20S subunit proteins in this cluster indicates a link between *ras* and the proteasome. *Ras*-activated kinases are known to regulate levels of other proteins (for example c-Myc (Sears, 2004) and retinoic acid receptors (Srinivas *et al.*, 2005)) via phosphorylation and induction of ubiquitin-proteasome protein degradation. *Ras* has also been shown to inhibit proteasome mediated degradation by promoting the stabilisation of otherwise degraded proteins, for example *ras*-induced stabilisation of p21 cyclin-dependent kinase inhibitor (Coleman *et al.*, 2003), however these effects are not thought to involve direct interaction between *ras* and the proteasome.

Recently, solute carrier family 4, member 11 has been shown to function as an Na/borate co-transporter involved in cell cycle control (Romero, 2005), raising the possibility that solute carrier family 4, member 1 represented in the *ras* profile cluster may also be involved in cell cycle control, possibly mediated by *ras*.

Interestingly, three profiles in the *ras* profile cluster represent a total of 60 eukaryotic proteins that have no known function. The domain families that generated these three profiles (shown in table 5.1) are the protein tyrosine phosphatase domain (26 proteins), fibronectin type 3 domain (9 proteins), and the NAD(P)-binding Rossmann-like domain (25 proteins). Since protein tyrosine phosphatase domains may be regulated by, or themselves regulate, *ras* (Zhang *et al.*, 2004), and cellular attachment to extracellular fibronectin type 3 domains has been shown be abolished by *ras* (Fang *et al.*, 1994), the 60 proteins of unknown function represented by these three profiles make intriguing targets for experimental study to determine the nature of any functional association to *ras*.

**Table 5.1 Profiles in the Ras Query Cluster.** *The 17 profiles that clustered with the user-defined Ras query profiles.*

| Depth of Profile | CATH Domain Code | # Relatives in Profile | Functional Description |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Homologous Superfamily | 1.20.1050.20 | 109 | Signal transduction/transcription activation |
| Homologous Superfamily | 2.60.40.630 | 104 | Signal transduction/transcription activation |
| Homologous Superfamily | 1.10.532.10 | 93 | Signal transduction/transcription activation |
| s30 | 1.20.58.30 | 47 | Solute carrier family 4, member 1 |
| s30 | 1.20.920.10 | 60 | Bromodomain: transcription regulation |
| s30 | 3.40.50.2020 | 23 | Uridine kinase |
| s30 | 3.40.50.720 | 25 | NAD(P)-binding Rossmann-like domain |
| s30 | 3.90.190.10 | 26 | Protein tyrosine phosphatase domain |
| s30 | 3.90.830.10 | 20 | Vesicle transport related |
| s30 | 1.10.10.60 | 25 | Homeobox: transcriptional regulation |
| s30 | 1.10.183.10 | 60 | Myosin heavy chain |
| s35 | 2.60.40.30 | 9 | Fibronectin type 3 domain |
| s35 | 3.60.20.0 | 25 | Proteasome 20S subunit |
| s50 | 3.90.490.0 | 16 | p33 ING1 tumour suppressor-like |
| s50 | 3.30.40.10 | 19 | Polycomb: transcription regulation |
| s70 | 1.10.10.60 | 37 | Homeobox: transcription regulation |
| s70 | 1.10.10.60 | 88 | Homeobox, transcription regulation |

## 5.5 Summary

Phylogenetic profiles built from Gene3D data have been shown to identify size dependent and universal CATH homologous superfamilies in prokaryotic genomes. A subset of universal size-dependent CATH homologous superfamilies (just 9% of CATH homologous superfamilies in prokaryotes) has been shown to account for 56% of prokaryotic genome domain assignments. The relationship between genome size and domain occurrence of metabolic domain families and regulatory domain families within these universal size-dependent CATH homologous superfamilies appears to follow microeconomic principles that may underlie determination of the size of prokaryotic genomes.

A novel protocol using Gene3D domain family subcluster phylogenetic occurrence profiles has been designed and shown to identify novel functionally associated domain clusters across complete genomes. The use of domain family subcluster occurrence profiles allows functional associations between specific domain family subclusters to be identified. Initial analysis of these clusters shows encouraging results. The method has been shown to identify known functionally associated proteins, as well as novel functionally associated proteins. These profile clusters can also be searched for significant matches to user-defined occurrence profiles, in which the user can specify the species distribution and genome occurrence in their own query profile. Initial analysis using user-defined *ras* (a well established signalling cascade protein) occurrence profiles identifies known *ras*-associated signal transduction and transcription regulators. In addition, three novel profiles representing a total of 60 functionally uncharacterised eukaryotic proteins have been identified as functionally associated with *ras*, and would make good experimental targets for further analysis.

# CHAPTER SIX

# Discussion and Future Work

## 6.1 Discussion

Gene3D is the first domain architecture database which characterises completely sequenced genomes by clustering into protein families and then assigning structural and sequence domain families from well characterised resources (CATH, Pfam). Throughout this thesis this resource is extensively described and used for genomic analysis to identify evolutionary relationships between individual components of completely sequenced genomes.

The identification of power law like behaviour in the size distributions of protein families, domain families and domain architecture families underlines the evolutionary relationships of proteins and protein domains in genomes, where few families dominate genome-space. The re-use of protein domains to form novel domain architectures in different protein families provides an illustration of the evolutionary strategy applied by Nature in genome evolution, whereby 50% of domain family assignments are common to all three Kingdoms of life, whilst only 16% of protein families are common to all three Kingdoms of life.

The power law distributions described in Gene3D can be exploited in several ways. Analysis of domain family distribution across multiple genomes has been used to estimate the number of folds in Nature. Structural genomics initiatives have already exploited Gene3D to identify novel domain families and prioritise structural genomics targets. Novel fold groups and structural families identified by analysing Gene3D data can be prioritised by several criteria including species distribution and fold group size, according to the strategic requirements of the different structural genomics consortia. Given the limited resources of structural genomics initiatives, it is vital to provide the best possible strategy for mapping fold space. The expansion of certain domain families in the genomes illustrates the bias in current protein structure classifications, where some domain families appear to be highly under-represented. Identification of

these expanded domain families can be used to select further targets to allow homology modelling of a greater proportion of the genome sequences.

The expansion of CATH domain families with sequences from completed genomes significantly increases the amount of functional data associated with these families. This data revealed those families in which multiple functions had evolved and highlighted the importance of considering domain context when inheriting functional properties between domain relatives. Analysis using Gene3D therefore allows domains in different contexts having different functional properties to be targeted for structure determination to reveal the structural mechanisms by which function evolves.

Analysis also identified 154,965 genscan sequences belonging to protein families containing known proteins, indicating that at least 57% of these genscan predictions are likely to be real protein sequences, thus warranting further study. This shows that gene predictions in eukaryotic genomes can be assessed for their reliability using protein family and domain assignment data.

The relationship between domain family frequency and genome size in bacteria was analysed. The identification of 85 universal, size-dependent domain families that are strongly correlated with bacterial genome size, representing just 9% of domain families but accounting for 56% of domain assignments in bacterial genomes highlights the dominance of a small number of domain families in bacterial genome evolution. This data adheres to microeconomics principles that can provide an explanation of the correlation between domain family usage and genome size in prokaryotes. The balance between the selection pressures of reproductive efficiency promoting small genomes, and environmental response and exploitation capacity promoting larger genomes, acts to enforce a balance between metabolic and regulatory genes a highly competitive system.

To attempt to understand the complex relationships between domain distribution across genomes and protein functional networks, a novel phylogenetic profile method for identifying functionally associated clusters of proteins, based upon domain family subcluster occurrence profiles was developed. This protocol enables identification of protein clusters sharing discrete domain distributions across genomes, which are indicative of functional and evolutionary associations. In addition to identification of clusters containing proteins with previously reported functional associations, novel

clusters containing proteins with no previously reported functional association were also identified. User-defined occurrence profiles have been used to search Gene3D profiles for proteins with functional associations to the user-defined query profile. Since these user-defined query profiles can be built according to the users specific requirements (in terms of species distribution and domain/protein occurrence identification method), they provide a readily accessible and easily interrogated resource. Initial analysis using *ras* query profiles (a well established signal transduction cascade protein) identified proteins known to be involved in ras-associated systems, such as signal transduction and transcription regulation. Additionally, 60 eukaryotic proteins with presently unknown functions were identified that are likely to have a functional association with *ras*. These proteins would make good targets for further study.

## 6.2 Future Work

Expansion of the Gene3D resource to include all completely sequenced genomes as they are completed would provide a more comprehensive resource for biologists interested in particular genomes. Addition of novel archaea and eukaryota genomes will permit a more rigorous analysis of these Kingdoms. Development of an efficient protocol for the automated inclusion of these additional genomes would therefore be of benefit.

As protein sequence databases increase in size, they produce an increase in the sensitivity of profile based methods of domain assignment. Additionally, source databases (for example Pfam and CATH) are updated; this necessitates regular updating of Gene3D data not only to incorporate expansions to domain family classifications, but also to reflect increases in performance of profile based methods. Development of efficient, fully automated protocols for updating domain assignments would be extremely beneficial to the resource.

The recent addition of Gene3D to the InterPro resource (which integrates major protein family classifications and provides regular mappings from major resources (UniProt, PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR SuperFamily and SUPERFAMILY)) permits Gene3D families to be integrated with definitions from other InterPro member databases and provides integration between InterPro domains and the CATH classification. Gene3D domain mappings can thus be supplemented with domain mappings from InterPro to further increase genome coverage and perhaps better define domain architectures, by including resources such as SCOP and TIGRFAMs. In addition, building HMMs from the largest Newfam domain families identified using Gene3D, would also contribute to increasing genome coverage and enable additional complete domain architectures to be identified. The use of recent profile-profile HMMs methods for the identification of distant relatives, which have been shown to be more sensitive than the HMM methods used in this thesis, would also increase domain assignment genome coverage, and hence allow a greater characterisation of completed genomes.

A program to compare domain strings using dynamic programming would be advantageous in identifying protein families having similar domain architectures, and

identifying similar domain architectures that have been defined using different domain family resources. This program could also help to identify protein families in Gene3D that contain multiple domain architectures. Although low in number, untangling these protein families may produce insight into the evolution of domain architectures. In order to facilitate identification of specific domain architecture families, HMMs representing complete domain architectures and partial domain architectures that are highly recurrent in the genomes should be developed. These HMMs would not only allow rapid classification of new genomes, but can be used to untangle protein families containing poorly defined domain architectures.

Further analysis of domain family subcluster phylogenetic profile correlations may permit identification of the mechanisms underlying the evolution of domain networks in bacterial genomes, and also identify biologically and medicinally important clusters. Inclusion of Gene3D derived data, such as domain occurrence profile cluster information, and phylogenetic profiles derived from protein families and domain architecture superfamilies, as well as addition of protein-protein interaction data would make the Gene3D resource more useful to biologists. However, it is important that the database and user interface is capable of rapid retrieval of user query results. The addition of genomes, domain assignments, additional functional resources and data derived from Gene3D analysis will only be of benefit if this data can still be accessed quickly.

With the recent advances in the field of bioinformatics, it might be wise, when considering the massive amount of different associations gleaned from current genomic data, to remember that currently we can only see a very tiny fraction of genome space. Only by exploring this genome space further can we hope to fully comprehend the myriad of evolutionary processes involved in the struggle for life.

# APPENDICES

**Appendix I – Genome Coverage in Gene3D.** *Coverage shown for each genome in Gene3D, identified by NCBI Taxonomy taxon identifier and organism name. Coverage is calculated as percentage of total proteins/residues in genome assigned only a CATH domain (Class 1-4), overlapping CATH and Pfam domains, and only a Pfam domain. The percentage of total annotated proteins/residues is also shown.*

| Taxon Id | Organism | Kingdom | % Coverage by Total Proteins | | | | % Coverage by Total Residues | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | CATH | CATH & Pfam | Pfam | Total Annotated | CATH | CATH & Pfam | Pfam | Total Annotated |
| 56636 | Aeropyrum pernix | A | 36.4 | 1.5 | 14.8 | **52.7** | 31.7 | 0.0 | 11.2 | **42.9** |
| 2234 | Archaeoglobus fulgidus | A | 41.4 | 2.8 | 17.3 | **61.5** | 36.6 | 0.0 | 14.0 | **50.6** |
| 64091 | Halobacterium sp. NRC-1 | A | 37.4 | 2.6 | 15.6 | **55.6** | 32.2 | 0.0 | 12.1 | **44.4** |
| 2190 | Methanocaldococcus jannaschii | A | 40.0 | 2.4 | 19.2 | **61.5** | 33.1 | 0.0 | 14.3 | **47.5** |
| 190192 | Methanopyrus kandleri AV19 | A | 35.4 | 3.7 | 26.9 | **66.0** | 28.8 | 0.0 | 19.7 | **48.5** |
| 188937 | Methanosarcina acetivorans C2A | A | 36.2 | 3.9 | 24.0 | **64.1** | 31.3 | 0.0 | 18.3 | **49.7** |
| 192952 | Methanosarcina mazei Goe1 | A | 36.0 | 4.3 | 27.4 | **67.7** | 31.3 | 0.0 | 20.7 | **52.1** |
| 187420 | Methanothermobacter thermautotroph. | A | 42.3 | 2.3 | 17.2 | **61.9** | 35.4 | 0.0 | 14.1 | **49.6** |
| 13773 | Pyrobaculum aerophilum | A | 32.0 | 2.6 | 22.5 | **57.0** | 30.7 | 0.0 | 17.9 | **48.6** |
| 29292 | Pyrococcus abyssi | A | 43.3 | 2.9 | 21.8 | **68.0** | 35.3 | 0.0 | 15.7 | **51.0** |
| 186497 | Pyrococcus furiosus DSM 3638 | A | 38.7 | 4.2 | 30.0 | **72.9** | 34.9 | 0.0 | 23.1 | **58.1** |
| 53953 | Pyrococcus horikoshii | A | 39.3 | 2.6 | 20.7 | **62.6** | 32.9 | 0.0 | 15.1 | **48.0** |
| 2287 | Sulfolobus solfataricus | A | 36.1 | 2.5 | 30.7 | **69.3** | 32.1 | 0.0 | 24.4 | **56.5** |
| 111955 | Sulfolobus tokodaii | A | 34.5 | 2.5 | 28.4 | **65.4** | 31.8 | 0.0 | 22.6 | **54.4** |
| 2303 | Thermoplasma acidophilum | A | 46.5 | 2.0 | 19.4 | **67.9** | 38.4 | 0.0 | 16.0 | **54.4** |
| 50339 | Thermoplasma volcanium | A | 45.5 | 1.9 | 18.9 | **66.3** | 38.1 | 0.0 | 15.6 | **53.7** |
| 181661 | Agrobacterium tumefaciens C58 Cereon | B | 45.5 | 5.1 | 28.6 | **79.1** | 37.8 | 0.0 | 22.6 | **60.4** |
| 180835 | Agrobacterium tumefaciens C58 U.Wash | B | 44.9 | 4.7 | 28.2 | **77.7** | 38.3 | 0.0 | 22.9 | **61.2** |
| 63363 | Aquifex aeolicus | B | 48.4 | 3.1 | 19.8 | **71.3** | 38.5 | 0.0 | 14.1 | **52.6** |
| 86665 | Bacillus halodurans | B | 42.7 | 2.4 | 16.1 | **61.1** | 36.2 | 0.0 | 13.4 | **49.6** |
| 1423 | Bacillus subtilis | B | 42.6 | 2.6 | 16.3 | **61.4** | 36.7 | 0.0 | 13.9 | **50.6** |
| 206672 | Bifidobacterium longum NCC2705 | B | 44.2 | 5.3 | 27.9 | **77.3** | 34.3 | 0.1 | 19.9 | **54.2** |
| 139 | Borrelia burgdorferi | B | 26.3 | 1.2 | 18.3 | **45.8** | 24.7 | 0.0 | 15.2 | **39.8** |
| 375 | Bradyrhizobium japonicum | B | 43.0 | 3.8 | 25.5 | **72.3** | 34.6 | 0.0 | 20.6 | **55.2** |
| 29459 | Brucella melitensis | B | 43.8 | 5.0 | 27.9 | **76.7** | 38.4 | 0.1 | 23.1 | **61.6** |
| 204722 | Brucella suis 1330 | B | 41.1 | 4.6 | 26.9 | **72.6** | 38.5 | 0.1 | 23.2 | **61.7** |
| 135842 | Buchnera aphidicola | B | 58.1 | 8.7 | 30.2 | **97.0** | 51.3 | 0.2 | 23.4 | **74.9** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 198804 | Buchnera aphidicola str. Sg | B | 56.7 | 8.8 | 31.4 | 96.9 | 50.2 | 0.2 | 24.7 | 75.1 |
| 107806 | Buchnera sp. APS | B | 63.8 | 3.1 | 16.6 | 83.4 | 50.8 | 0.0 | 10.4 | 61.2 |
| 197 | Campylobacter jejuni | B | 45.5 | 2.6 | 18.1 | 66.1 | 35.8 | 0.0 | 13.6 | 49.4 |
| 190650 | Caulobacter crescentus CB15 | B | 44.8 | 4.6 | 25.5 | 75.0 | 39.0 | 0.1 | 20.2 | 59.2 |
| 83560 | Chlamydia muridarum | B | 43.6 | 2.2 | 18.7 | 64.4 | 33.0 | 0.0 | 12.6 | 45.6 |
| 813 | Chlamydia trachomatis | B | 42.9 | 2.7 | 20.2 | 65.8 | 33.0 | 0.0 | 14.5 | 47.5 |
| 115711 | Chlamydophila pneumoniae AR39 | B | 36.1 | 1.9 | 18.3 | 56.2 | 29.5 | 0.0 | 13.7 | 43.2 |
| 115713 | Chlamydophila pneumoniae CWL029 | B | 38.1 | 2.0 | 19.4 | 59.5 | 29.7 | 0.0 | 13.8 | 43.5 |
| 138677 | Chlamydophila pneumoniae J138 | B | 37.6 | 2.0 | 19.2 | 58.7 | 29.4 | 0.0 | 13.7 | 43.0 |
| 194439 | Chlorobium tepidum TLS | B | 39.5 | 4.8 | 23.0 | 67.3 | 38.9 | 0.1 | 20.5 | 59.4 |
| 1488 | Clostridium acetobutylicum | B | 39.6 | 5.6 | 26.9 | 72.1 | 34.2 | 0.1 | 21.8 | 56.1 |
| 1502 | Clostridium perfringens | B | 39.0 | 6.6 | 29.5 | 75.1 | 34.3 | 0.1 | 24.1 | 58.5 |
| 212717 | Clostridium tetani E88 | B | 38.6 | 7.0 | 31.1 | 76.6 | 32.6 | 0.1 | 23.6 | 56.2 |
| 196164 | Corynebacterium efficiens YS-314 | B | 39.4 | 4.8 | 26.1 | 70.3 | 32.6 | 0.0 | 20.4 | 53.0 |
| 196627 | Corynebacterium glutamicum | B | 41.9 | 4.3 | 24.9 | 71.1 | 34.8 | 0.0 | 20.2 | 55.1 |
| 1299 | Deinococcus radiodurans | B | 42.9 | 2.1 | 14.4 | 59.4 | 34.4 | 0.0 | 10.2 | 44.6 |
| 199310 | Escherichia coli CFT073 | B | 36.0 | 4.6 | 30.5 | 71.1 | 34.2 | 0.1 | 27.3 | 61.6 |
| 83333 | Escherichia coli K12 | B | 45.4 | 2.7 | 19.3 | 67.4 | 37.1 | 0.0 | 15.9 | 53.0 |
| 83334 | Escherichia coli O157:H7 | B | 37.5 | 4.0 | 28.7 | 70.3 | 32.9 | 0.0 | 23.7 | 56.6 |
| 155864 | Escherichia coli O157:H7 EDL933 | B | 39.7 | 2.3 | 19.3 | 61.3 | 33.2 | 0.0 | 15.6 | 48.8 |
| 190304 | Fusobacterium nucleatum | B | 37.1 | 5.3 | 27.5 | 69.9 | 32.7 | 0.1 | 23.1 | 55.8 |
| 71421 | Haemophilus influenzae Rd | B | 49.7 | 2.9 | 20.0 | 72.5 | 41.5 | 0.0 | 14.9 | 56.5 |
| 85962 | Helicobacter pylori 26695 | B | 38.6 | 2.4 | 19.2 | 60.2 | 31.0 | 0.0 | 15.1 | 46.1 |
| 85963 | Helicobacter pylori J99 | B | 40.7 | 2.5 | 20.6 | 63.8 | 31.6 | 0.0 | 15.6 | 47.3 |
| 220668 | Lactobacillus plantarum WCFS1 | B | 42.7 | 5.7 | 27.9 | 76.3 | 36.4 | 0.1 | 23.9 | 60.3 |
| 1360 | Lactococcus lactis subsp. lactis | B | 46.8 | 2.2 | 16.2 | 65.2 | 37.8 | 0.0 | 12.5 | 50.4 |
| 189518 | Leptospira interrogans serovar lai 56601 | B | 27.4 | 3.2 | 17.3 | 48.0 | 27.6 | 0.0 | 16.6 | 44.2 |
| 1642 | Listeria innocua | B | 41.7 | 5.5 | 28.7 | 75.9 | 35.7 | 0.1 | 23.4 | 59.2 |
| 169963 | Listeria monocytogenes EGD-e | B | 44.4 | 6.1 | 29.6 | 80.1 | 37.9 | 0.1 | 24.5 | 62.4 |
| 381 | Mesorhizobium loti | B | 45.1 | 1.9 | 14.3 | 61.3 | 36.6 | 0.0 | 11.6 | 48.2 |
| 1769 | Mycobacterium leprae | B | 48.3 | 2.4 | 14.0 | 64.6 | 40.7 | 0.0 | 9.5 | 50.2 |
| 83331 | Mycobacterium tuberculosis CDC1551 | B | 38.1 | 4.2 | 25.9 | 68.2 | 34.2 | 0.0 | 19.1 | 53.3 |
| 83332 | Mycobacterium tuberculosis H37Rv | B | 44.3 | 2.0 | 16.1 | 62.4 | 34.7 | 0.0 | 11.0 | 45.8 |
| 2097 | Mycoplasma genitalium | B | 50.0 | 3.9 | 18.6 | 72.5 | 34.9 | 0.0 | 10.4 | 45.3 |
| 28227 | Mycoplasma penetrans | B | 36.8 | 5.7 | 18.1 | 60.7 | 25.1 | 0.0 | 14.1 | 39.3 |
| 2104 | Mycoplasma pneumoniae | B | 38.6 | 3.2 | 22.4 | 64.2 | 27.8 | 0.1 | 12.2 | 40.1 |
| 2107 | Mycoplasma pulmonis | B | 37.2 | 6.8 | 26.1 | 70.1 | 28.3 | 0.1 | 18.4 | 46.8 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 122586 | Neisseria meningitidis MC58 | B | 38.1 | 4.9 | 28.6 | 71.7 | 35.5 | 0.1 | 25.1 | 60.6 |
| 122587 | Neisseria meningitidis Z2491 | B | 37.9 | 4.7 | 28.6 | 71.2 | 35.7 | 0.1 | 25.4 | 61.2 |
| 103690 | Nostoc sp. PCC 7120 | B | 34.6 | 5.4 | 23.1 | 63.0 | 31.5 | 0.1 | 17.7 | 49.2 |
| 182710 | Oceanobacillus iheyensis | B | 42.0 | 5.3 | 28.7 | 75.9 | 38.4 | 0.1 | 24.3 | 62.7 |
| 747 | Pasteurella multocida | B | 49.7 | 3.0 | 20.0 | 72.7 | 39.5 | 0.0 | 14.2 | 53.7 |
| 208964 | Pseudomonas aeruginosa PAO1 | B | 47.7 | 2.8 | 17.7 | 68.1 | 37.0 | 0.0 | 14.3 | 51.3 |
| 160488 | Pseudomonas putida KT2440 | B | 43.2 | 5.3 | 29.8 | 78.3 | 36.0 | 0.0 | 24.6 | 60.6 |
| 305 | Ralstonia solanacearum | B | 42.2 | 4.3 | 27.1 | 73.7 | 33.9 | 0.0 | 22.0 | 56.0 |
| 781 | Rickettsia conorii | B | 29.4 | 4.2 | 23.3 | 56.9 | 31.3 | 0.1 | 22.5 | 53.9 |
| 782 | Rickettsia prowazekii | B | 47.4 | 3.1 | 21.1 | 71.6 | 36.7 | 0.0 | 14.8 | 51.5 |
| 90370 | Salmonella enterica | B | 37.4 | 5.1 | 32.0 | 74.5 | 33.8 | 0.1 | 28.1 | 62.0 |
| 99287 | Salmonella typhimurium LT2 | B | 40.7 | 5.6 | 34.7 | 81.1 | 35.4 | 0.1 | 29.6 | 65.0 |
| 211586 | Shewanella oneidensis MR-1 | B | 37.9 | 5.7 | 28.6 | 72.2 | 32.8 | 0.0 | 23.8 | 56.6 |
| 198214 | Shigella flexneri 2a str. 301 | B | 41.0 | 4.9 | 37.9 | 83.8 | 36.0 | 0.1 | 29.6 | 65.6 |
| 382 | Sinorhizobium meliloti | B | 45.8 | 4.7 | 28.2 | 78.8 | 38.5 | 0.0 | 23.0 | 61.6 |
| 158878 | Staphylococcus aureus Mu50 | B | 40.0 | 5.7 | 29.5 | 75.2 | 36.0 | 0.1 | 25.4 | 61.5 |
| 196620 | Staphylococcus aureus MW2 | B | 40.3 | 5.3 | 30.3 | 75.9 | 36.2 | 0.1 | 25.3 | 61.6 |
| 158879 | Staphylococcus aureus N315 | B | 42.1 | 5.9 | 29.9 | 77.9 | 37.1 | 0.1 | 25.5 | 62.7 |
| 176280 | Staphylococcus epidermidis | B | 40.4 | 5.2 | 27.7 | 73.3 | 37.8 | 0.1 | 24.1 | 62.0 |
| 208435 | Streptococcus agalactiae 2603V/R | B | 41.4 | 5.6 | 28.5 | 75.6 | 37.1 | 0.1 | 23.1 | 60.2 |
| 211110 | Streptococcus agalactiae NEM316 | B | 42.5 | 6.1 | 27.2 | 75.7 | 36.3 | 0.1 | 21.3 | 57.7 |
| 210007 | Streptococcus mutans UA159 | B | 44.0 | 6.1 | 26.3 | 76.5 | 39.6 | 0.1 | 21.8 | 61.5 |
| 171101 | Streptococcus pneumoniae R6 | B | 41.3 | 6.3 | 26.9 | 74.4 | 38.2 | 0.1 | 22.9 | 61.2 |
| 170187 | Streptococcus pneumoniae TIGR4 | B | 40.2 | 6.0 | 27.4 | 73.5 | 37.8 | 0.1 | 23.5 | 61.4 |
| 160490 | Streptococcus pyogenes M1 GAS | B | 42.9 | 6.5 | 28.8 | 78.1 | 37.4 | 0.1 | 23.2 | 60.6 |
| 186103 | Streptococcus pyogenes MGAS8232 | B | 40.8 | 6.0 | 27.3 | 74.1 | 36.5 | 0.1 | 23.3 | 59.9 |
| 198543 | Streptococcus pyogenes phage 315.6 | B | 40.0 | 5.8 | 27.6 | 73.5 | 35.8 | 0.1 | 23.4 | 59.3 |
| 100226 | Streptomyces coelicolor A3(2) | B | 41.8 | 4.3 | 23.9 | 70.1 | 33.5 | 0.0 | 18.3 | 51.8 |
| 1148 | Synechocystis sp. PCC 6803 | B | 43.5 | 3.0 | 15.3 | 61.8 | 34.2 | 0.0 | 10.5 | 44.8 |
| 119072 | Thermoanaerobacter tengcongensis | B | 39.2 | 6.7 | 28.7 | 74.6 | 35.7 | 0.1 | 22.9 | 58.7 |
| 197221 | Thermosynechococcus elongatus BP-1 | B | 40.2 | 6.0 | 26.3 | 72.5 | 35.8 | 0.1 | 21.0 | 56.9 |
| 2336 | Thermotoga maritima | B | 45.6 | 2.9 | 20.1 | 68.6 | 36.8 | 0.0 | 14.4 | 51.2 |
| 160 | Treponema pallidum | B | 40.1 | 1.9 | 15.2 | 57.1 | 29.7 | 0.0 | 10.1 | 39.8 |
| 2130 | Ureaplasma urealyticum | B | 40.6 | 2.4 | 14.0 | 57.0 | 26.1 | 0.0 | 7.7 | 33.8 |
| 666 | Vibrio cholerae | B | 41.1 | 2.8 | 17.5 | 61.4 | 34.1 | 0.0 | 13.5 | 47.6 |
| 164609 | Wigglesworthia brevipalpis | B | 52.1 | 7.3 | 31.2 | 90.7 | 48.2 | 0.2 | 24.7 | 73.1 |
| 190486 | Xanthomonas axonopodis | B | 42.2 | 5.2 | 26.9 | 74.3 | 35.6 | 0.0 | 20.9 | 56.6 |
| 190485 | Xanthomonas campestris | B | 42.3 | 5.4 | 27.9 | 75.5 | 35.8 | 0.0 | 21.5 | 57.4 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 160492 | Xylella fastidiosa 9a5c | B | 33.9 | 2.2 | 13.2 | **49.3** | 33.2 | 0.0 | 11.3 | **44.6** |
| 183190 | Xylella fastidiosa Temecula1 | B | 40.2 | 6.3 | 28.0 | **74.5** | 36.0 | 0.1 | 21.9 | **58.0** |
| 632 | Yersinia pestis | B | 40.1 | 5.2 | 34.0 | **79.3** | 34.6 | 0.1 | 27.2 | **61.9** |
| 187410 | Yersinia pestis KIM | B | 38.4 | 5.1 | 32.0 | **75.4** | 34.2 | 0.1 | 26.4 | **60.6** |
| 7165 | Anopheles gambiae | E | 40.8 | 7.0 | 19.9 | **67.8** | 26.4 | 0.0 | 14.3 | **40.8** |
| 3702 | Arabidopsis thaliana | E | 40.8 | 5.9 | 26.1 | **72.9** | 27.1 | 0.0 | 16.2 | **43.4** |
| 6239 | Caenorhabditis elegans | E | 33.0 | 5.7 | 26.2 | **65.0** | 21.8 | 0.0 | 17.3 | **39.2** |
| 7955 | Danio rerio | E | 42.1 | 5.6 | 11.6 | **59.2** | 37.0 | 0.0 | 11.7 | **48.7** |
| 7227 | Drosophila melanogaster | E | 39.9 | 6.7 | 21.4 | **67.9** | 22.7 | 0.0 | 13.8 | **36.5** |
| 6035 | Encephalitozoon cuniculi | E | 36.4 | 3.3 | 18.6 | **58.4** | 24.0 | 0.0 | 12.1 | **36.2** |
| 55529 | Guillardia theta polymorph | E | 39.8 | 4.3 | 23.2 | **67.2** | 32.3 | 0.1 | 14.9 | **47.3** |
| 9606 | Homo sapiens | E | 41.8 | 9.0 | 19.3 | **70.2** | 27.9 | 0.0 | 14.2 | **42.1** |
| 10090 | Mus musculus | E | 45.6 | 10.2 | 18.5 | **74.3** | 33.5 | 0.0 | 14.6 | **48.1** |
| 36329 | Plasmodium falciparum 3D7 | E | 28.3 | 3.3 | 20.8 | **52.5** | 9.8 | 0.0 | 11.4 | **21.2** |
| 10116 | Rattus norvegicus | E | 35.2 | 5.4 | 13.1 | **53.7** | 24.1 | 0.0 | 9.7 | **33.9** |
| 4932 | Saccharomyces cerevisiae | E | 36.9 | 3.6 | 16.5 | **57.0** | 21.6 | 0.0 | 11.1 | **32.7** |
| 4896 | Schizosaccharomyces pombe | E | 40.5 | 5.9 | 27.2 | **73.7** | 25.7 | 0.0 | 17.4 | **43.1** |
| 31033 | Takifugu rubripes | E | 32.9 | 5.3 | 13.5 | **51.7** | 30.6 | 0.0 | 14.4 | **45.1** |

**Appendix II – Universal Domain Families in Gene3D.** *The 212 universal CATH and Pfam domain families, where each domain family is universal to all three Kingdoms in Gene3D. Universal domain families are found to occur in a minimum of 70% of the genomes from a Kingdom, so these domain families are found in at least 70% of the genomes of each Kingdom in Archaea, Bacteria and Eukaryota. Domain families denoted by their CATH or Pfam identification are shown with the percentage of genomes in each Kingdom in which they are identified: %A(rchaea), %B(acteria) and %E(ukaryota) and the most common GO function associated with the domain family.*

| Domain Family | %A | %B | %E | GO Function |
|---|---|---|---|---|
| 1.10.10.10 | 100 | 98 | 100 | regulation of transcription; DNA-dependent;transcription factor activity |
| 1.10.10.250 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 1.10.10.60 | 100 | 81 | 100 | regulation of transcription; DNA-dependent;transcription factor activity |
| 1.10.1030.10 | 81 | 82 | 79 | carbamoyl-phosphate synthase activity;ATP binding;cytoplasm |
| 1.10.1060.10 | 100 | 83 | 79 | electron transport;electron transporter activity |
| 1.10.1140.10 | 100 | 93 | 93 | ATP-binding and phosphorylation-dependent chloride channel activity |
| 1.10.1160.10 | 88 | 99 | 86 | glutamate-tRNA ligase activity;ATP binding;glutamyl-tRNA aminoacylation |
| 1.10.15.10 | 100 | 94 | 86 | base-excision repair;DNA binding |
| 1.10.150.30 | 81 | 99 | 79 | intracellular;DNA binding |
| 1.10.260.10 | 100 | 88 | 86 | DNA binding;regulation of transcription; DNA-dependent |
| 1.10.260.30 | 100 | 99 | 71 | RNA binding;GTP binding;signal recognition particle |
| 1.10.275.10 | 100 | 93 | 79 | catalytic activity;lyase activity |
| 1.10.287.10 | 100 | 99 | 93 | protein biosynthesis;structural constituent of ribosome |
| 1.10.287.310 | 100 | 99 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 1.10.287.40 | 81 | 100 | 71 | tRNA ligase activity;serine-tRNA ligase activity;ATP binding |
| 1.10.290.10 | 100 | 98 | 86 | DNA topoisomerase type I activity;nucleic acid binding |
| 1.10.340.10 | 100 | 96 | 86 | base-excision repair;DNA binding |
| 1.10.40.30 | 94 | 92 | 71 | catalytic activity;lyase activity |
| 1.10.455.10 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 1.10.460.10 | 100 | 98 | 86 | nucleic acid binding;DNA topoisomerase type I activity;DNA modification |
| 1.10.560.10 | 100 | 98 | 93 | chaperone activity;ATP binding;protein folding |
| 1.10.600.10 | 100 | 93 | 86 | isoprenoid biosynthesis;transferase activity |
| 1.10.730.10 | 100 | 100 | 86 | tRNA ligase activity;ATP binding; |
| 1.10.8.50 | 94 | 100 | 93 | RNA binding;structural constituent of ribosome;intracellular |
| 1.10.8.60 | 100 | 100 | 100 | ATP binding;nucleotide binding |
| 1.20.1010.10 | 100 | 100 | 86 | ATP binding;arginine-tRNA ligase activity |
| 1.20.120.140 | 94 | 99 | 86 | GTP binding;signal recognition particle;protein targeting |
| 1.20.200.10 | 100 | 93 | 79 | catalytic activity;lyase activity |
| 1.20.58.100 | 88 | 73 | 79 | electron transport;oxidoreductase activity; |
| 1.25.40.10 | 81 | 94 | 100 | nucleus;intracellular |
| 2.10.230.10 | 75 | 100 | 86 | chaperone activity;protein folding; |
| 2.130.10.10 | 88 | 84 | 100 | nucleus;membrane |
| 2.160.10.10 | 88 | 94 | 79 | transferase activity;acyltransferase activity |
| 2.170.120.12 | 94 | 100 | 93 | DNA binding;DNA-directed RNA polymerase activity;transcription |
| 2.20.29.10 | 100 | 100 | 86 | translation elongation factor activity;GTP binding;translational elongation |
| 2.30.30.30 | 100 | 99 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 2.30.35.20 | 100 | 100 | 86 | structural constituent of ribosome;intracellular;ribosome |
| 2.30.35.30 | 100 | 94 | 79 | ligase activity;ATP binding |
| 2.30.42.10 | 94 | 94 | 100 | protein binding;proteolysis and peptidolysis |
| 2.40.10.80 | 94 | 93 | 93 | ATP-binding and phosphorylation-dependent chloride channel activity |

| | | | | |
|---|---|---|---|---|
| 2.40.150.20 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 2.40.240.10 | 100 | 80 | 86 | protein biosynthesis;structural constituent of ribosome |
| 2.40.30.10 | 100 | 100 | 93 | GTP binding;translation elongation factor activity |
| 2.40.33.10 | 81 | 92 | 86 | pyruvate kinase activity;glycolysis; |
| 2.40.40.30 | 100 | 99 | 93 | DNA binding;DNA-directed RNA polymerase activity;nucleus |
| 2.40.50.100 | 94 | 99 | 86 | membrane;protein secretion |
| 2.40.50.140 | 100 | 100 | 93 | protein biosynthesis;RNA binding |
| 2.40.50.150 | 100 | 99 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 2.70.150.10 | 81 | 86 | 93 | ATP binding;membrane |
| 2.70.20.10 | 81 | 96 | 79 | DNA binding;DNA topoisomerase type I activity;DNA topological change |
| 2.70.40.10 | 100 | 87 | 86 | dUTP metabolism;hydrolase activity |
| 3.10.129.10 | 88 | 93 | 79 | catalytic activity;metabolism |
| 3.10.180.10 | 81 | 81 | 79 | lactoylglutathione lyase activity;carbohydrate metabolism |
| 3.10.20.30 | 88 | 93 | 93 | electron transporter activity;electron transport |
| 3.10.20.70 | 100 | 74 | 71 | glutamate-ammonia ligase activity;nitrogen fixation |
| 3.10.290.10 | 100 | 100 | 93 | RNA binding;pseudouridylate synthase activity |
| 3.10.50.40 | 88 | 98 | 93 | protein folding;isomerase activity |
| 3.20.19.10 | 100 | 72 | 79 | metabolism;lyase activity |
| 3.20.20.100 | 81 | 72 | 93 | oxidoreductase activity;electron transporter activity |
| 3.20.20.105 | 100 | 87 | 71 | queuine tRNA-ribosyltransferase activity;queuosine biosynthesis |
| 3.20.20.120 | 100 | 98 | 86 | catalytic activity;metabolism |
| 3.20.20.140 | 100 | 100 | 86 | hydrolase acting on carbon-nitrogen (but not peptide) bonds; in cyclic amides |
| 3.20.20.150 | 100 | 83 | 71 | endonuclease activity;DNA repair |
| 3.20.20.170 | 81 | 100 | 79 | lyase activity;fructose-bisphosphate aldolase activity |
| 3.20.20.60 | 100 | 100 | 86 | kinase activity;transferase activity |
| 3.20.20.70 | 100 | 92 | 86 | lyase activity;carbohydrate metabolism |
| 3.20.20.90 | 100 | 100 | 93 | oxidoreductase activity;electron transport |
| 3.30.160.30 | 100 | 98 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 3.30.190.20 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 3.30.200.20 | 88 | 90 | 100 | ATP binding;protein amino acid phosphorylation |
| 3.30.230.10 | 100 | 100 | 100 | protein biosynthesis;ATP binding |
| 3.30.230.20 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 3.30.300.20 | 100 | 100 | 79 | nucleic acid binding;RNA binding |
| 3.30.360.10 | 75 | 89 | 93 | oxidoreductase activity;electron transport |
| 3.30.390.10 | 100 | 98 | 86 | catalytic activity;metabolism |
| 3.30.390.30 | 100 | 94 | 79 | electron transport;oxidoreductase activity |
| 3.30.420.10 | 100 | 100 | 93 | DNA binding;DNA recombination |
| 3.30.420.100 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 3.30.420.40 | 81 | 100 | 100 | ATP binding;heat shock protein activity |
| 3.30.420.80 | 100 | 99 | 93 | protein biosynthesis;structural constituent of ribosome |
| 3.30.428.10 | 94 | 99 | 86 | transferase activity;UTP-hexose-1-phosphate uridylyltransferase activity |
| 3.30.470.20 | 100 | 97 | 79 | ligase activity;ATP binding |
| 3.30.499.10 | 100 | 72 | 79 | metabolism;lyase activity |
| 3.30.540.10 | 88 | 88 | 71 | inositol/phosphatidylinositol phosphatase activity;hydrolase activity |
| 3.30.550.10 | 100 | 94 | 79 | glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity |
| 3.30.56.20 | 94 | 97 | 86 | phenylalanine-tRNA ligase activity;tRNA ligase activity |
| 3.30.565.10 | 100 | 100 | 100 | ATP binding;kinase activity |
| 3.30.70.100 | 100 | 81 | 71 | metal ion transport;metal ion binding |
| 3.30.70.141 | 100 | 80 | 86 | nucleoside-diphosphate kinase activity;ATP binding;GTP biosynthesis |
| 3.30.70.160 | 100 | 100 | 86 | transaminase activity;transferase activity |
| 3.30.70.20 | 100 | 77 | 93 | electron transport;electron transporter activity |
| 3.30.70.210 | 100 | 92 | 86 | nucleic acid binding;RNA binding |
| 3.30.70.240 | 100 | 100 | 93 | translation elongation factor activity;GTP binding;translational elongation |
| 3.30.70.330 | 100 | 98 | 100 | nucleic acid binding;RNA binding |

| | | | | |
|---|---|---|---|---|
| 3.30.70.350 | 100 | 100 | 86 | structural constituent of ribosome;intracellular;ribosome |
| 3.30.70.460 | 100 | 83 | 71 | catalytic activity;ligase activity |
| 3.30.70.530 | 100 | 100 | 93 | DNA binding;DNA-directed RNA polymerase activity;transcription |
| 3.30.70.580 | 81 | 94 | 86 | pseudouridylate synthase activity;tRNA processing |
| 3.30.70.60 | 100 | 98 | 86 | protein biosynthesis;structural constituent of ribosome |
| 3.30.70.600 | 100 | 97 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 3.30.70.660 | 81 | 96 | 86 | pseudouridylate synthase activity;tRNA processing |
| 3.30.70.780 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 3.30.70.810 | 88 | 98 | 79 | arginine-tRNA ligase activity;ATP binding;arginyl-tRNA aminoacylation |
| 3.30.860.10 | 100 | 96 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 3.30.930.10 | 100 | 100 | 86 | tRNA ligase activity;ATP binding |
| 3.40.1010.10 | 100 | 94 | 86 | metabolism;methyltransferase activity |
| 3.40.1060.10 | 100 | 70 | 79 | metabolism;lyase activity |
| 3.40.120.10 | 100 | 98 | 86 | carbohydrate metabolism; phosphotransferases |
| 3.40.190.10 | 100 | 99 | 93 | transport;transporter activity |
| 3.40.190.80 | 88 | 87 | 71 | inositol/phosphatidylinositol phosphatase activity;hydrolase activity |
| 3.40.225.10 | 88 | 70 | 79 | isomerase activity;lyase activity |
| 3.40.250.10 | 75 | 92 | 86 | transferase activity;thiosulfate sulfurtransferase activity |
| 3.40.30.10 | 94 | 100 | 93 | electron transport;electron transporter activity |
| 3.40.309.10 | 81 | 84 | 71 | oxidoreductase activity;metabolism |
| 3.40.350.10 | 81 | 79 | 71 | proteolysis and peptidolysis;metalloexopeptidase activity |
| 3.40.367.20 | 100 | 87 | 93 | structural constituent of cytoskeleton;actin cytoskeleton |
| 3.40.430.10 | 75 | 94 | 86 | 5-amino-6-(5-phosphoribosylamino)uracil reductase activity |
| 3.40.440.10 | 94 | 84 | 71 | GTP binding;purine nucleotide biosynthesis |
| 3.40.460.10 | 100 | 98 | 86 | carbohydrate metabolism; phosphotransferases |
| 3.40.47.10 | 100 | 93 | 86 | transferase activity;fatty acid biosynthesis |
| 3.40.470.10 | 81 | 98 | 86 | DNA repair;uracil DNA N-glycosylase activity |
| 3.40.50.1000 | 100 | 100 | 93 | hydrolase activity;metabolism |
| 3.40.50.10050 | 100 | 99 | 93 | GTP binding;translation elongation factor activity |
| 3.40.50.1010 | 100 | 100 | 86 | DNA binding;nuclease activity |
| 3.40.50.1050 | 100 | 92 | 71 | carbohydrate metabolism; phosphotransferases |
| 3.40.50.1090 | 100 | 76 | 79 | intracellular;transcription factor activity |
| 3.40.50.1100 | 100 | 84 | 71 | amino acid metabolism;lyase activity |
| 3.40.50.1140 | 88 | 83 | 86 | electron transport;oxidoreductase activity |
| 3.40.50.1220 | 100 | 86 | 93 | regulation of transcription; DNA-dependent;electron transport |
| 3.40.50.1260 | 100 | 98 | 86 | phosphoglycerate kinase activity;glycolysis |
| 3.40.50.1270 | 100 | 98 | 86 | phosphoglycerate kinase activity;glycolysis |
| 3.40.50.1370 | 100 | 88 | 79 | amino acid metabolism; carboxyl- and carbamoyltransferase activity |
| 3.40.50.140 | 100 | 98 | 79 | nucleic acid binding;DNA modification |
| 3.40.50.1400 | 75 | 80 | 79 | ferrochelatase activity;heme biosynthesis |
| 3.40.50.1440 | 75 | 93 | 93 | GTP binding;structural molecule activity |
| 3.40.50.150 | 100 | 100 | 100 | S-adenosylmethionine-dependent methyltransferase activity |
| 3.40.50.1580 | 100 | 98 | 79 | nucleoside metabolism;catalytic activity |
| 3.40.50.1820 | 94 | 98 | 86 | catalytic activity;hydrolase activity |
| 3.40.50.1900 | 100 | 87 | 71 | lyase activity;amino acid metabolism |
| 3.40.50.1940 | 100 | 96 | 86 | sugar binding;carbohydrate metabolism |
| 3.40.50.1950 | 100 | 89 | 71 | lyase activity;carboxy-lyase activity |
| 3.40.50.1990 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 3.40.50.20 | 94 | 94 | 79 | ligase activity;ATP binding |
| 3.40.50.2000 | 81 | 94 | 79 | transferase activity; transferring hexosyl groups;metabolism |
| 3.40.50.2020 | 100 | 98 | 86 | nucleoside metabolism;transferase activity |
| 3.40.50.300 | 100 | 100 | 100 | ATP binding;nucleotide binding |
| 3.40.50.50 | 81 | 91 | 86 | pyruvate kinase activity;glycolysis |
| 3.40.50.610 | 100 | 100 | 93 | ATP binding;ligase activity |

| | | | | |
|---|---|---|---|---|
| 3.40.50.620 | 100 | 100 | 93 | nucleotidyltransferase activity;biosynthesis |
| 3.40.50.720 | 100 | 100 | 100 | oxidoreductase activity;metabolism |
| 3.40.50.7700 | 94 | 79 | 71 | phosphoribosylaminoimidazole carboxylase activity;'de novo' IMP biosynthesis |
| 3.40.50.790 | 100 | 100 | 86 | structural constituent of ribosome;intracellular;ribosome |
| 3.40.50.800 | 100 | 100 | 86 | tRNA ligase activity;ATP binding |
| 3.40.50.850 | 88 | 72 | 79 | catalytic activity;metabolism |
| 3.40.50.880 | 100 | 98 | 86 | catalytic activity;glutamine metabolism |
| 3.40.50.920 | 100 | 99 | 86 | oxidoreductase activity;electron transport |
| 3.40.50.970 | 100 | 99 | 93 | oxidoreductase activity;metabolism |
| 3.40.510.10 | 100 | 100 | 86 | ATP binding;tRNA ligase activity |
| 3.40.605.10 | 81 | 86 | 71 | oxidoreductase activity;metabolism |
| 3.40.630.10 | 100 | 99 | 86 | proteolysis and peptidolysis;metallopeptidase activity |
| 3.40.630.30 | 100 | 94 | 100 | N-acetyltransferase activity;transferase activity |
| 3.40.640.10 | 100 | 100 | 86 | transaminase activity;transferase activity |
| 3.40.718.10 | 100 | 73 | 79 | oxidoreductase activity;metabolism |
| 3.50.30.20 | 81 | 82 | 71 | catalytic activity;carbamoyl-phosphate synthase activity;ATP binding |
| 3.50.50.60 | 100 | 100 | 86 | electron transport;oxidoreductase activity |
| 3.50.7.10 | 100 | 98 | 93 | chaperone activity;ATP binding |
| 3.60.15.10 | 100 | 100 | 93 | hydrolase activity;molecular_function unknown |
| 3.60.20.10 | 100 | 94 | 100 | metabolism;endopeptidase activity |
| 3.60.21.10 | 100 | 92 | 93 | hydrolase activity;protein serine/threonine phosphatase activity |
| 3.90.110.10 | 81 | 86 | 79 | oxidoreductase activity;L-lactate dehydrogenase activity |
| 3.90.170.10 | 88 | 84 | 79 | GTP binding;purine nucleotide biosynthesis |
| 3.90.180.10 | 81 | 78 | 79 | alcohol dehydrogenase activity; zinc-dependent;zinc ion binding |
| 3.90.188.10 | 88 | 91 | 79 | ribonucleoside-diphosphate reductase activity; DNA replication |
| 3.90.226.10 | 94 | 94 | 93 | catalytic activity;metabolism |
| 3.90.230.10 | 100 | 100 | 93 | proteolysis and peptidolysis;metalloexopeptidase activity |
| 3.90.244.10 | 88 | 98 | 86 | ribonucleoside-diphosphate reductase activity; DNA replication |
| 3.90.269.10 | 100 | 82 | 79 | glutamate-ammonia ligase activity;nitrogen fixation |
| 3.90.470.10 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 3.90.550.10 | 100 | 100 | 86 | transferase activity;nucleotidyltransferase activity |
| 3.90.700.10 | 88 | 78 | 79 | oxidoreductase activity;electron transport |
| 3.90.740.10 | 100 | 100 | 86 | tRNA ligase activity;ATP binding;amino acid activation |
| 3.90.77.20 | 100 | 92 | 86 | kinase activity;transferase activity |
| 3.90.79.10 | 100 | 92 | 86 | hydrolase activity;isoprenoid biosynthesis |
| 3.90.80.10 | 81 | 73 | 86 | metabolism;membrane;pyrophosphatase activity |
| 3.90.800.10 | 88 | 96 | 86 | glutamate-tRNA ligase activity;ATP binding;glutamyl-tRNA aminoacylation |
| 3.90.870.10 | 100 | 98 | 79 | 3;4 dihydroxy-2-butanone-4-phosphate synthase activity |
| 3.90.930.12 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| 4.10.910.10 | 100 | 100 | 93 | RNA binding;structural constituent of ribosome;intracellular |
| 4.10.950.10 | 100 | 99 | 93 | structural constituent of ribosome;intracellular;ribosome |
| PF00083 | 100 | 99 | 93 | transport;membrane |
| PF00324 | 81 | 88 | 86 | amino acid-polyamine transporter activity;amino acid transport;membrane |
| PF00344 | 100 | 100 | 93 | protein secretion;protein translocase activity;membrane |
| PF00534 | 100 | 91 | 86 | biosynthesis;transferase activity |
| PF00571 | 100 | 99 | 79 | membrane;transport |
| PF00572 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| PF00573 | 100 | 99 | 93 | structural constituent of ribosome;intracellular;ribosome |
| PF00588 | 75 | 97 | 79 | RNA binding;RNA processing;RNA methyltransferase activity |
| PF00673 | 100 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| PF00999 | 100 | 79 | 93 | regulation of pH;integral to membrane |
| PF01066 | 100 | 96 | 86 | phospholipid biosynthesis;transferase activity |
| PF01192 | 100 | 77 | 93 | DNA-directed RNA polymerase activity;transcription; DNA-dependent |
| PF01509 | 100 | 92 | 86 | pseudouridylate synthase activity;RNA processing |

| | | | | |
|---|---|---|---|---|
| PF01513 | 100 | 89 | 79 | kinase activity;transferase activity |
| PF01725 | 100 | 88 | 86 | hydrolase activity;molecular_function unknown |
| PF01842 | 75 | 78 | 71 | metabolism;amino acid binding |
| PF01966 | 100 | 93 | 79 | catalytic activity;hydrolase activity |
| PF01979 | 100 | 78 | 71 | hydrolase activity;N-acetylglucosamine-6-phosphate deacetylase activity |
| PF02272 | 100 | 100 | 71 | nucleic acid binding;ATP binding |
| PF03946 | 88 | 100 | 93 | structural constituent of ribosome;intracellular;ribosome |
| PF04560 | 100 | 100 | 93 | DNA binding;DNA-directed RNA polymerase activity;transcription |
| PF04561 | 100 | 90 | 93 | DNA-directed RNA polymerase activity;transcription |
| PF05362 | 75 | 76 | 79 | ATP-dependent peptidase activity;serine-type endopeptidase activity |

**Appendix III – Universal Size-Dependent Superfamilies in Bacteria.** *The 66 universal size-dependant superfamilies used in this analysis are denoted by their CATH code and PDB representative. The total number of domain assignments, percentage of genomes containing these assignments and Spearman's Rank correlation coefficient between domain occurrence and genome size is shown. The size distribution group (Pl – power law, L – linear, Log – logarithmic) and functional classification (R –regulatory, M – metabolic, O – other, P – poorly characterised) are indicated.*

| CATH Code | PDB Rep. | Number of Domains | Universality | Spearmans Rank Coefficient | Size Distiribution Function | Func. Code | Functional Definition |
|---|---|---|---|---|---|---|---|
| 1.10.10.10 | 1lea00 | 5695 | 98 | 0.951 | Pl | R | Winged helix (DNA-binding) |
| 3.40.190.10 | 1anf02 | 4595 | 99 | 0.867 | Pl | R | SBP-bacterial 1 periplasmic binding protein |
| 1.10.10.60 | 1mbe00 | 3391 | 86 | 0.926 | Pl | R | Homeodomain-like (DNA-binding) |
| 3.40.50.2600 | 8abp02 | 3101 | 85 | 0.865 | Pl | R | CheY receiver domain |
| 3.30.565.10 | 1cuk03 | 2400 | 100 | 0.861 | Pl | R | Histidine kinase |
| 3.40.50.1820 | 1dqzA0 | 2338 | 97 | 0.878 | Pl | M/P | Alpha and Beta hydrolases |
| 1.10.260.10 | 1neq00 | 1684 | 89 | 0.773 | Pl | R | l-repressor (DNA-binding) |
| 2.40.50.100 | 1htp00 | 1056 | 98 | 0.742 | Pl | M | Biotin-requiring enzymes |
| 3.10.180.10 | 1kw3B1 | 778 | 82 | 0.823 | Pl | M | Glyoxalase/bleomycin/dioxygenase |
| 3.90.180.10 | 1qorA1 | 737 | 79 | 0.851 | Pl | M | Zinc-binding dehydrogenase |
| 3.40.605.10 | 1ag8A1 | 663 | 85 | 0.879 | Pl | M | NADP-oxidoreductase |
| 3.10.20.30 | 4fxc00 | 637 | 92 | 0.721 | Pl | R | TGS-domain (nuclotide bindingin regulation) |
| 2.130.10.10 | 2bbkH0 | 555 | 84 | 0.76 | Pl | R | WD domain (Beta-transduction) |
| 3.10.129.10 | 1mkaA0 | 535 | 93 | 0.824 | Pl | M | Thioesterase/Maoc-like domain |
| 3.40.50.1420 | 1mjhA0 | 456 | 80 | 0.75 | Pl | R | Universal stress protein |
| 3.40.50.1090 | 1cf9A3 | 323 | 79 | 0.843 | Pl | R | DJ-1-Pfpl transcription regulator |
| 3.20.20.120 | 1oneA2 | 320 | 99 | 0.746 | Pl | M | Enolase Ct-domain/methylaspartate ammonia-lyase |
| 3.30.70.130 | 2chsA0 | 241 | 77 | 0.787 | Pl | R | Endoribonuclease |
| 3.40.50.850 | 1nbaA0 | 234 | 75 | 0.77 | Pl | M | Isochorismatase family |
| 3.30.43.10 | 1uxy02 | 172 | 80 | 0.7 | Pl | M | FAD-binding domain |
| 3.40.50.300 | 1efuA1 | 18,292 | 100 | 0.958 | L | M | P-loop containing nucleotide triphosphate |
| 3.40.50.720 | 1evyA1 | 6880 | 100 | 0.939 | L | M | NAD(P)-binding domain |
| 3.40.50.150 | 1admA1 | 3844 | 100 | 0.87 | L | M | Methyltransferase |
| 3.50.50.60 | 3ladA2 | 3101 | 100 | 0.91 | L | M | NADH-FAD oxidoreductases |
| 3.40.640.10 | 1tplA2 | 2498 | 100 | 0.917 | L | M | Type 1 PLP-dependent aspartate aminotransferase |
| 3.40.30.10 | 1abe00 | 2098 | 99 | 0.86 | L | M | Thioredoxin-like domain |
| 3.40.630.30 | 1cjwA0 | 1814 | 96 | 0.868 | L | M/R | Acetyltransferase |
| 3.90.550.10 | 1qg8A0 | 1805 | 100 | 0.801 | L | O | SpsA-glycosyltransferase |
| 3.30.420.10 | 1rthA5 | 1635 | 100 | 0.72 | L | O | RNAase-H |
| 3.40.50.970 | 1poxA3 | 1611 | 100 | 0.8 | L | M | Thiamine diphosphate binding fold |
| 3.40.50.980 | 1lcl01 | 1583 | 81 | 0.79 | L | M | Acetyl-CoA synthetase family |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3.40.47.10 | 1pxtA1 | 1118 | 94 | 0.804 | L | M | Beta-oxidation,lipid metabolism |
| 3.90.226.10 | 1dubA1 | 871 | 94 | 0.758 | L | M | Enoyl-CoA hydratase |
| 3.40.710.10 | 2bltA0 | 842 | 85 | 0.768 | L | O | Beta-lactamases |
| 3.40.630.10 | 2ctc00 | 819 | 99 | 0.808 | L | M | Zinc-metallopeptidase domain |
| 3.90.77.20 | 1rkd02 | 719 | 94 | 0.784 | L | M | Carbohydrate kinase |
| 1.10.443.10 | 1aihA0 | 714 | 93 | 0.734 | L | O | Phage integrase domain |
| 2.160.10.10 | 1lxa01 | 647 | 94 | 0.797 | L | M | Acetyltransferase |
| 2.30.42.10 | 1pdr00 | 495 | 94 | 0.705 | L | R | PDZ domain (signaling protein) |
| 3.40.50.1900 | 1qoqB2 | 491 | 89 | 0.767 | L | M | Amino acid metabolism |
| 3.40.250.10 | 1rhs01 | 488 | 89 | 0.791 | L | M | Rhodanase domain |
| 3.60.20.10 | 1gdoA0 | 478 | 95 | 0.736 | L | M | Glutamine amidotransferase class-II |
| 3.20.20.100 | 1ads00 | 434 | 73 | 0.795 | L | M | Aldo/ketoreductases |
| 3.20.20.60 | 1pkm02 | 430 | 100 | 0.771 | L | M | Pyruvate kinases |
| 3.20.20.140 | 1a4mA0 | 417 | 100 | 0.762 | L | M | Metal-dependent hydrolases |
| 3.40.50.1140 | 1aa8A1 | 398 | 83 | 0.812 | L | M | FAD-dependent oxidoreductase |
| 2.40.40.20 | 1eu1A4 | 378 | 72 | 0.706 | L | M | VAT-N domain (binding protein) |
| 3.30.499.10 | 1c96A3 | 364 | 76 | 0.714 | L | M | Acotinase |
| 3.20.20.10 | 1bd0A1 | 356 | 83 | 0.716 | L | M | Alanine racemase/pyridoxal binding |
| 1.20.200.10 | 1fupA2 | 343 | 95 | 0.739 | L | M | Lyase |
| 3.90.230.10 | 1chmA2 | 338 | 100 | 0.72 | L | M/O | M24 metallopeptidase family |
| 3.40.109.10 | 1nox00 | 297 | 83 | 0.755 | L | M | NAD-NADPH oxidoreductase |
| 3.40.50.170 | 1garA0 | 241 | 90 | 0.723 | L | M | Formyltransferase |
| 3.30.1090.10 | 1qdlA2 | 199 | 84 | 0.714 | L | M | Chorismate binding enzyme |
| 3.40.718.10 | 1iso00 | 179 | 77 | 0.76 | L | M | Isocitrate/isopropyl malatede hydrogenase |
| 3.90.269.10 | 1lgr01 | 171 | 85 | 0.772 | L | M | Glutamine synthetase |
| 3.20.20.20 | 1ad4B0 | 153 | 90 | 0.748 | L | M | Pterin binding enzyme (methyltransferases) |
| 3.30.470.10 | 3daaA1 | 146 | 77 | 0.75 | L | M | Aminotransferases class IV |
| 2.40.50.140 | 1ckmA2 | 2292 | 100 | 0.709 | Log | O | RNA-binding domain |
| 3.20.20.90 | 1tpfA0 | 2049 | 100 | 0.898 | Log | M | FMN-dependent enzymes |
| 3.40.50.1000 | 1jud01 | 1642 | 100 | 0.838 | Log | M | Dehalogenase |
| 3.30.470.20 | 1low02 | 1150 | 97 | 0.894 | Log | M | ATP-grasp fold |
| 3.40.50.880 | 1gpmA1 | 840 | 98 | 0.751 | Log | M | Glutamine amidotransferase class I |
| 3.90.79.10 | 1mut00 | 729 | 94 | 0.889 | Log | O | DNA repair domain |
| 3.40.50.620 | 1dnpA1 | 675 | 100 | 0.712 | Log | O | DNA repair |
| 3.60.21.10 | 4kbpA2 | 560 | 94 | 0.723 | Log | R/O | Calcineurin-like phosphoesterase |

# REFERENCES


Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol.* Oct 5;215(3):403-410

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* Sep 1;25(17):3389-3402

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. and Murzin, A. G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* Jan 1;32:D226-229

Anscombe, J. F. (1973) Graphs in statistical analysis. *Am Stat.* 27, 17-21

Apic, G., Gough, J. and Teichamnn, S. A. (2001) Domain combinations in archaeal, eubacterial and eularyotic proteomes. *J Mol Biol.* Jul 6;310(2):311-325

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.* 25:25-29

Atkinson, A. C. (1985) Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis. In (Copas, J. B. *et al.*, eds). Oxford University Press.

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, A. K., Taylor, P., Uddin, A. and Zygouri, C. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* 31:400-402

Avery, O. T., MacLeod, C. and McCarty, M. (1944) Studies on the chamical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental Medicine* 79(2):137-158

Babu, M. M. and Teichmann, S. A. (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res.* 31:1234-1244

Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28:45-48

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. and Yeh, L. S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* Jan 1;33(database issue):D154-159

Baldessari, D., Shin, Y., Krebs, O., Konig, R., Koide, T., Vinayagam, A., Fenger, U., Mochii, M., Terasaka, C., Kitayama, A., Peiffer, D., Ueno, N., Eils, R., Cho, K. W. and Niehrs, C. (2005) Global gene expression profiling and cluster analysis in Xenopus laevis. *Mech. Dev.* Mar;122(3):441-475.

Bansal, a. K. (1999) An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics* Nov 15(11):900-908

Barbazuk, W. B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., Bedell, J. A., McPherson, J. D. and Johnson, S. L. (2000) The syntenic relationship of the zebrafish and human genomes. *Genome Res.* Sept 10(9):1351-1358

Barrett, T., Suzek, T., Troup, D., Wilhite, S., Ngau, W., Ledoux, P., Rudnev, D., Lash, A., Fujibuchi, W. and Edger, R. (2005) NCBI GEO: mining millions of expression profiles - database and tools. *Nucleic Acids Res.* 33(datbase issue):D562-D566

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J>, Yeats, C. and Eddy, S. R. (2004) The Pfam protein families database. *Nucleic Acids Res.* Jan 1(32):D138-141

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. (2000) GenBank. *Nucleic Acids Res.* Jan 1;28(database issue):D15-18

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L. (2005) GenBank. *Nucleic Acids Res.* Jan 1;33(database issue):D34-38

Bernal, A. Ear, U. and Kyrpides, N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* Jan 1;29(1):126-127

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol.* May 25 112(3):535-542

Bird, A. P. (1995) Gene number, noise reduction and biological complexity. *Trends Genet.* 11:94-100

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M. (2003) The SWISSPROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* Jan 1;31(1):365-370

Bork, P. and Koonin, E. V. (1998) Predicting functions from protein sequences – where are the bottlenecks? *Nature Genet.* Apr 18(4):313-318

Borodovsky, M. and McIninch, J. (1993) Recognition of genes in DNA sequence with ambiguities. *Biosystems* 30(1-3):161-171

Bray, J. E., Todd, A. E., Pearl, F. M., Thornton, J. M. and Orengo, C. A. (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng.* Mar;13(3):153-165

Bryant, S. H. and Lawrence, C. E. (1993) An empirical energy function for threading protein sequence through the folding motif. *Proteins* May 16(1):92-112

Buchan, D. W., Rison, S. C., Bray, J. E., Lee, D., Pearl, F. M., Thornton, J. M. and Orengo, C. A. (2003) Gene3D: structural assignments for the biologist and bioinformaticist alike. *Nucleic Acids Res.* Jan 1;31(1):469-473

Buchan, D. W., Shepherd, A. J., Lee, D., Pearl, F. M., Rison, S. C., Thornton, J. M. and Orengo, C. A. (2002) Gene3D: structural assignments for whole genes and genomes using the CATH domain structure database. *Genome Res.* Mar;12(3):503-514

Burge, C. and Karlin, S. (1997) Prediction of complete gene structure in human genomic DNA. *J Mol Biol.* Apr 25 268(1):78-94

Campbell, J. A., Davies, G. J., Bulone, V. and Henrissat, B. (1998) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem J.* Sep 15;326:929-939

Chandonia, J. M. and Brenner, S. E. (2005) Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* Jan1;58(1):166-179

Chen, J., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S., Hurwitz, D. I., Jackson, J. D., Jacobs, A. R., Lanczycki, C. J., Liebert, C. A., Liu, C., Madej, T., Marchler-Bauer, A., Marchler, G. H., Mazumder, R., Nikolskaya, A. N., Rao, B. S., Panchenko, A. R., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Vasudevan, S., Wang, Y., Yamashita, R. A., Yin, J. J. and Bryant, S. H. (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.* Jan 1;31(1):474-477

Cherkasov, A. and Jones, S. J. (2004) Structural characterization of genomes by large scale sequence-structure threading. *BMC Bioinformatics*. 2004 Apr 3;5:37

Chothia, C. (1992) Proteins. One thousand families for the molecular biologist. *Nature* 357:543-544

Chothia, C., Gough, J., Vogel, C. and Teichmann, S. A. (2003) Evolution of the protein repertoire. *Science* 300:1701-1703

Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* Apr;5(4):823-6

Coleman, M. L., Marshall, C. J. and Olson, M. F. (2003) Ras promotes p21 (Waf1/Cip1) protein stability via a cyclin D1-imposed block in proteasome-mediated degradation. *EMBO J.* May 1;22(9):2036-2046.

Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z. and Speed, T. P. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 20(3):323-331

Copley, R. R. and Bork, P. (2000) Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J Mol Biol* 303:627-641

Coulson, A. F. W. and Moult, J. (2002) A Unifold, Mesofold and Superfold Model of Protein Fold Use. *Proteins* 46:61-71

Dandekar, T., Snel, B., Huynen, m. and Borl, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *TIBS* Sept 23:324-328

Date, S. and Marcotte, E. M. (2003) Discovery of uncharacterised cellular systems by genome-wide analysis of functional linkages. *Nature Biotech.* 21(9):1055- 1062

Dayhoff, M. O. (ed) 1965-1978. Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington.

Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978) A model of evolutionary change in proteins. In: Atlas of protein sequence and structure 5(3) M. O.

de Bruijn, F. J. et al. (1998) Bacterial Genomes: Physical Structure and Analysis. Structure and Sizes of Genomes of the Archaea and Bacteria. Kluwer Academic Publishers.

Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Fend, Z., Green, R. K., Flippen-Anderson, J. L., Westbrook, J., Berman, H. M. and Bourne, P. E. (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.* Jan 1;33(database issue):D233-237

Devos, S. and Valencia, A. (2000) Practical limits of function prediction. *Proteins: Structure, Function, and Genetics*.41:98-107

Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M. and Holm, L. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.* Jan 1;29(1):55-7

Dobrindt, U. and Hacker, J. (2001) Whole-genome plasticity in pathogenic bacteria. *Curr Opin Microbiol.* Oct;4(5):550-557

Doolittle, R. F. (1990) Searching through sequence databases. *Methods Enzymol.* 183:99-110

Doolittle, R. F. (1995) The multiplicity of domains in proteins. *Annu Rev Biochem.* 64:287-314

Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* 14:755-763

Eisenberg, D., Marcotte, E. M., Xenarios, I. and Yeates, T. O. (2000) Protein function in the post-genomic era. *Nature* June 405:823-826

Elofsson, A. and Sonnhammer, E. L. (1999) A compairson of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics* Jun;15(6):480-500

Enright, A. J. and Ouzounis, C. A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* May 16(5):451- 457

Enright, A. J., Kunin, V. and Ouzounis, C. A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* Aug 1;31(15):4632-4638

Enright, A.J., van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30:1575-1584

Fang, K. S., Barker, K., Sudol, M. and Hanafusa, H. (1994) A transmembrane protein-tyrosine phosphatase contains spectrin-like repeats in its extracellular domain. *J Biol Chem.* 1994 May 13;269(19):14056-63.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van DEN Berghe, A., Volckaert, G. and Ysebaert, M. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* Apr 8 260(5551):500-507

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G. G., FitzHugh, W., Fields, C. A., Gocayne, J. D., Scott, J. D., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. and Venter, J. C. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223):496-512

Fleming, K., Muller, A., MacCallum, R. M. and Sternberg, M. J. (2004) 3D-GENOMICS: a database to compare structural and functional annotations of proteins between sequenced genomes. *Nucleic Acids Res.* Jan 1;32:D245-250

Franklin, R. E. and Gosling, R. G. (1953) Molecular configuration in sodium thymonucleate. *Nature* 171:740-741

Frizelle, G. (1998) The Management of Complexity in Manufacturing. Business Intelligence Press.

Gamov, G., Rich, A., Ycas, M. (1956) The problem of information transfer from the nucleic acids to proteins. *Adv Biol Med Phys.* 4:23-68

Gibrat, J. F., Madej, T. and Bryant, S. H. (1996) Surprising similarities in structure comparison. *Curr Opin Struct Biol.* Jun;6(3):377-385

Goldovsky, L., Cases, I., Enright, A. J. and Ouzounis, C. A. (2005) BioLayout(Java): Versatile Network Visualisation of Structural and Functional Relationships. Appl *Bioinformatics* 4(1):71-74

Goudreau, P. N. and Stock, A. M. (1998) Signal transduction in bacteria: molecular mechanisms of stimulus-response coupling. *Curr Opin Microbiol.* 1:160-169

Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol.* Nov 2;313(4):903-919

Govindarajan, S. and Goldstein, R. A. (1996) Why are some proteins structures so common? *Proc Natl Acad Sci* USA 93:3341-3345

Govindarajan, S., Recabarren, R. and Goldstein, R. A. (1999) Estimating the total number of protein folds. *Proteins* 35:408-414

Groft, C. M., Beckmann, R., Sali, A., Burley, s. K. (2000) Crystal structure of ribosome anti-association factor If6. *Nature Struct Biol.* 7:1156

Grant, A. D., Lee, D. and Orengo, C. (2004) Progress towards mapping the universe of protein folds. *Genome Biol.* 5(5):107

Guan, X. (1997) Domain identification by clustering sequence alignments. *Proc Int Conf Intell Syst Mol Biol.* 5:124-130

Haft, D.H., Selengut, J.D. and White, O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.* Jan 1;31(1):371-373

Harrison, A., Pearl, F. Sillitoe, I., Slidel, T., Mott, R., Thornton, J. and Orengo, C. (2003) Recognising the fold of a protein structure. *Bioinformatics* Sep 22;19(14):1748-1759

Harrison, A., Pearl, F., Mott, R., Thornton, J. and Orengo, C. (2002) Quantifying the similarities within fold space. *J Mol Biol.* Nov 8;323(5):909-926

Harrison, P. M. and Gerstein, M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J Mol Biol.* May 17;318(5):1155-1174

Heger, A. and Holm, L. (2003) Exhaustive Enumeration of Protein Domain Families. *J Mol Biol.* 328:749-767

Hegyi, H. and Gerstein, M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* Oct 11(10):1632-1640

Hegyi, H., Lin, J., Greenbaum, D. and Gerstein, M. (2002) Structural genomes analysis: characteristics of atypical, common, and horizontally transferred folds. *Proteins* 2:126-141

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K. and Hood, L. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science* Oct 24;278(5338):609-614

Hofmann, E., Wrench, P. M., Sharples, F. P., Hiller, R. G., Welte, W. and Diederichs, K. (1996) Structural basis of light harvesting by carotenoids: peridium-chlorophyll-protein from Amphidinium carterae. *Science* 272:1788

Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol.* Sep 5;233(1):123-138

Holm, L. and Sander, C. (1994) Parser for protein folding units. *Proteins* Jul;19(3):256-268

Holm, L. and Sander, C. (1996) The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.* Jan 1;24(1):206-209

Huang, H., Barker, W. C., Chen, Y. and Wu, C. H. (2003) iProClass: an integrated databse of protein family, function and structure information. *Nucleic Acids Res.* Jan 1;31(1):390-392

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C. and Birney, E. (2005) Ensembl s005. *Nucleic Acids Res.* Jan 1;33:D447-453

Jacobs, J. J., Kieboom, K., Marino, S., DePinho, R. A. and van Lohuizen, M. (1999) The oncogene and Polycomb-group gene bmi-1 regulates cell proliferation and senescence through the ink4a locus. *Nature* Jan 14;397(6715):164-168.

Jones, D. T., Taylor, W. R. And Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* Jul 2;358(6381):86-89

Jones, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C. and Thornton, J. M. (1998) Domain assignment for protein structures using a consensus approach: characterisation and analysis. *Protein Sci* 7:233-42

Jordan, I. K., Makarova, K. S., Spouge, J. L., Wolf, Y. I. and Koonin, E. V. (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* Apr;11(4):555-565

Kanehisa, M., Goto, S., Kawahima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32(database issue):D277-D280

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F. G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M. A., Tzouvara, K., Vaughn, R., Wu, D., Zhu, W. and Apweiler, R. (2005) The EMBL nucleotide sequence database. *Nucleic Acids Res.* Jan 1:33(database issue):D29-33

Kaplan, N. and Linial, M. (2005) Automatic detection of false annotations via binary property clustering. *BMC Bioinformatics* Mar 8;6(1):46

Kaplan, N., Sasson, O., Inbar, U., Friedlich, M., Fromer, M., Fleischer, H., Portugaly, E., Linial, N. and Linial M. (2005) ProtoNet 4.0: hierarchical classification of one million protein sequences. *Nucleic Acids Res.* Jan 1;33(1):D216-218

Karev, G. P., Wolf, Y. I. and Koonin, E. V. (2003) Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics* 19:1889-1900

Karev, G. P., Wolf, Y. I., Rzhetsky, A. Y., Berezovskaya, F. S. and Koonin, E. V. (2002) Birth and death of protein domains: A simple model of evolution explains power law behaviour. *BMC Evol Biol* 2: 8

Karp, P. D., Raley, S. and Zhu, J. (2001) Database verification studies of SWISS-PROT and GenBank. *Bioinformatics* 17(6):526-532.

Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14(10):846-856

Kendall, M. and Gibbons, J. D. (1990) Rank Correlation Methods (5th edition), Oxford University Press.

Kinch, L. N., Wrabl, J. O., Krishna, S. S., Majumdar, I., Sadreyev, R. I., Qi, Y., Pei, J., Cheng, H. and Grishin, N. V. (2003) CASP5 assessment of fold recognition target predictions. *Proteins* 53(6):395-409.

Klewer, D. A., Hoskins, A., Zhang, P., Davisson, V. J., Bergstrom, D. E., Liwang, A. C. (2000) NMR structure of a DNA duplex containing nucloeside analog 1-(2'-deoxy-beta-d-ribofuranosyl)-3-nitropyrrole and the structure of the unmodified control. *Nucleic Acids Res.* 28:4514

Kolesov, G., Mewes, H. W. and Frishman, d. (2001) SNAPping up functionally related genes based on context information: a colinearity-free approach. *J Mol Biol.* Aug 24;311(4):639-656

Koonin, E. V., Wolf, Y. I. and Karev, G. P. (2002) The structure of the protein universe and genome evolution. *Nature* 420:218-223

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modelling. *J Mol Biol.* Feb 235(5):1501-1531

Kunin, V., Cases, I., Enright, A. J., deLorenzo, V. and Ouzounis, C. A. (2003) Myriads of protein families, and still counting. *Genome Biol.* Jan 29;4(2):401

Kunin, V., Teichmann, S. A., Huynen, M. A. and Ouzounis, C. A. (2005) The properties of protein family spcae depend on experimental design. *Bioinformatics* Jun 1;21(11):2618-2622

Kurland, C. G., Canback, B. and Berg, O. G. (2003) Horizontal gene transfer: a critical view. *Proc Natl Acad Sci* USA Aug 19;100(17):9658-9662

Leder, P. and Nirenberg, M. W. (1964) RNA codewords and protein synthesis, 3. On the nucleotide sequence of a cysteine and a leucine DNA codeword. *PNAS* Dec 52:1521-1529

Lee, D., Grant, A., Marsden, R. L. and Orengo, C. (2005) Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins* May 15;59(3):603-615

Leonov, H., Mitchell, J. S. B. and Arkin, I. T. (2003) Monte Carlo Estimation of the Number of Possible Protein Folds: Effects of Sampling Bias and Folds Distributions. *Proteins* 51:352-359

Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P. and Bork, P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* Jan 1;30(1)242-244

Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. *Science* Mar 22;227(4693):1435-1441

Liu, J. and Rost, B. (2002) Target space for structural genomics revisited. *Bioinformatics* Jul;18(7):922-933

Liu, J. and Rost, B. (2003) Domains, motifs and clusters in the protein universe. *Curr Opin Chem Biol* 7:5-11

Liu, J. and Rost, B. (2004) CHOP proteins into structural domain-like fragments. *Proteins* 55(3):678-688

Luscombe, N. M., Qian, J., Zhang, Z., Johnson, T. and Gerstein, M. (2002) The domainance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol.* Jul 25;3(8):Epub RESEARCH0040

Maclean, J. A., Rao, M. K., Doyle, K. M., Richards, J. S. and Wilkinson, M. F. (2005) Regulation of the Rhox5 Homeobox Gene in Primary Granulosa Cells: Preovulatory Expression and Dependence on SP1/SP3 and GABP. *Biol Reprod.* Dec;73(6):1126-1134.

Madera, M. and Gough, J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res* 30(19):4321-4328

Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C. and Gough, J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.* Jan 1;32:D235-239

Mankiw, N. G. (1998) Principles of Microeconomics, 2nd edition. Dryden Press.

Maxam, A. M. and Gilbert, W. (1977) A new method for sequencing DNA. *PNAS* Feb 74(2):560-564

McGuffin, L. J. and Jones, D. T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* May 1;19(7):874-81

McGuffin, L. J., Street, S. A., Bryson, K., Sorensen, S. A. and Jones, D. T. (2004) The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucleic Acids Res.* Jan 1;32:D196-199

McShane, L. M., Radmacher, M, D., Freidlin, B., Yu, R., Li, M. and Simon, R. (2002) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* 18(11):1462-1469

Meinel, T., Krause, A., Luz, H., Vingron, M. and Staub, E. (2005) The SYSTERS Protein Family Database in 2005. *Nucleic Acids Res.* Jan 1;33:D226-229

Meinel, T., Vingron, M. and Krause, A. (2003) The SYSTERS Protein Family Database: taxon-related protein family size distributions and singleton frequencies. *Proc Ger Conf Bioinf.* Belleville, Munich, Germany, pp103-108

Mira, A., Ochman, H. and Moran, N. A. (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet.* Oct; 17(10):589-596

Mizuguchi, K., Deane, C. M., Blundell, T. L. and Overington, J. P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* Nov;7(11):2469-2471

Mobley, H. L. T., Island, M. D. and Hausinger, R. P. (1995) Molecular biology of microbial ureases. *Microbiological Rev.* Sep;59(3):451-480.

Moran, N. A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108:583-586

Mrowka, R. Patzak, A. and Herzel, H. (2001) Is there a bias in proteome research? *Genome Research* 11:1971-1973

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, >., Maslen, J., McDowall, J., Mitchell, A., Nilolskaya, A. N., Orchard, S.E., Pagni, M., Ponting, C. P., Quevillon, E., Selengut, J., Sigrist, C. J., Silventoinen, V., Studholme, D. J., Vaughan, R. and Wu, C. H. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.* Jan 1;(33):D201-205

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Hom, G. and Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* 51 Pt1:263-273

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* Apr 7;247(4):536-540

Nagano, N., Orengo, C. A. and Thornton, J. M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* 321:741-765

Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* Mar 48(3):443-453

Nirenberg, M. and Leder, P. RNA codewords and protein synthesis. The effect of trinucleotides upon the binding of sRNA to ribosomes. *Science* Sep 25(145):1399-1407

Nirenberg, M. W. and Matthaei, J. H. (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *PNAS* Oct 15(47):1588-1602

Ochman, H., Lawrence, J. G. and Groisman, E. A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* May 18;405(6784):299-304

Orengo, C. A., Sillitoe, I., Reeves, G. Pearl, F. M. (2001) Review: what can structural classifications reveal about protein evolution? *J Struct Biol.* 134:145-165.

Orengo, C. A. (1999) CORA - topological fingerprints for protein structural families. *Protein Sci.* Apr;8(4):699-715

Orengo, C. A. and Taylor, W. R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* 266:617-635

Orengo, C. A. and Thornton, J. M. (2005) Protein Families and their Evolution – A Structural Perspective. *Annu. Rev. Biochem.* 74:867-900

Orengo, C. A., Jones, D. T., Thornton, J. M. (1994) Protein superfamilies and domain superfolds. *Nature* 372:631-634

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. and Thornton, J. M. (1997) CATH - A hierarchical classification of protein domain structures. *Structure* 5:1093–1108

Orr, H. A. (2000) Adaptaion and the cost of complexity. *Evolution Int. J. Org. Evolution.* Feb 5;54:13-20

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *PNAS* 96:2896-2901

Pagel, P., Wong, P. and Frishman, D. (2004) A domain interaction map based on phylogenetic profiling. *J Mol Biol.* Dec 10;344(5):1331-1346

Papageorgiou, A. C., Shapiro, R. and Acharya, K. R. (1997) Molecular recognition of human angiogenin by placental ribonuclease inhibitor - an X-ray crystallographic study at 2.0 A resolution. *EMBO J* 16:5162

Park, J. and Teichmann, S. A. (1998) DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single and multi-domain proteins. *Bioinformatics* 14(2):144-150

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol.* 284:1201-1210

Patthy, L. (1999) Genome evoution and the evolution of exon-shuffling - a review. *Gene* Sep 30 238(1):103-114

Pavlidis, P. and Noble, W. S. (2001) Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology* 2(10):RESEARCH0042 Epub

Pawlowski, K., Rychlewski, L., Zhang, B. and Godzik, A. (2001) Fold predictions for Bacterial Genomes. *J Struct Biol.* 134:219-231

Pearl, F., Lee, D., Bray, J. E., Buchan, D. W., Shepherd, A., J. and Orengo, C. A. (2002) The CATH extended protein family database: providing structural annotations for genome sequences. *Protein Sci.* Feb;11(2):233-244

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J. and Orengo, C. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.* Jan 1;33:D247-251

Pellegrini, M., Marcotte, E. M., Thompson, M, J., Eisenberg, D. and Yeates, T. O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *PNAS* 96:4285-4288

Peloponese, J. M., Gregoire, C., Opi, S., Esquireu, D., Sturgis, J., Lebrun, E., Meurs, E., Collette, Y., Olive, D., Aubertin, A. M., Witvrow, M., Pannecouque, C., De Clercq, E., Bailly, C., Lebreton, J. and Loret, E. P. (2000) 1H-13C nuclear magnetic resonance assignment and structural characterisation of HIV-1 tat protein. *C.R.Acad. Sci.* III 323:883

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A and Baimeister, K. (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucloetides. *Science* Oct 16 238(4235):336-341

Pruitt K. D., Tatusova, T., Maglott, D. R. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence datbase of genomes, transcripts and proteins. *Nucleic Acids Res.* Jan 1;33(database issue):D501-504

Qian, J., Luscombe, N. M. and Gerstein, M. (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol* 313: 673-681

Ranea *et al.*, manuscript submitted.

Reese, M. G., Hartzell, G., Harris, N. L., Ohler, U., Abril, J. F. and Lewis, S. E. (2000) Genome annotation assessment in Drosophila melanogaster. *Genome Res.* Apr 10(4):483-501

Reeves *et al.*, manuscript in preparation.

Romero, M. F. (2005) Molecular pathophysiology of SLC4 bicarbonate transporters. *Curr Opin Nephrol Hypertens.* Sep;14(5):495-501.

Rossmann, M. G. and Argos, P. (1976) Exploring the structural homology of proteins. *J Mol Biol.* Jul 25;105(1):75-95

Rost, B. (2002) Enzyme function less conserved than anticipated. *J Mol Biol.* Apr 26;318(2):595-608

Rost, B. and Liu, J. (2003) The PredictProtein server. *Nucleic Acids Res* 31:3300- 3304

Rufino, S. D. and Blundell, T. L. (1994) Structure-based identification and clustering of protein families and superfamilies. *J Comput Aided Mol Des.* Feb 8(1):5-27

Russell, R. B. and Barton, G. J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* Oct 14(2):309-323

Saito, R., Suzuki, H. and Hayashizaki, Y. (2003) Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics* 19(6):756-763

Sali, A. and Blundell, T. L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simmulated annealing and dynamic programming. *J Mol Biol.* Mar 20;212(2):403-428

Salzberg, S. L., Delcher, A. L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* Jan 15 26(2):544-548

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., Peterson, G. B. (1982) Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol.* Dec 25;162(4):729-73

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *PNAS* Dec 74(12):5463-5467

Sasson, O., Vaaknin, A., Fleischer, H., Portugaly, E. Bilu, Y., Linial, N. and Linial M. (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res.* Jan 1;31(1):348-352

Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. and Altschul, S. F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* Dec 15(12):1000-1011

Schwartz, D. C. and Cantor, C. R. (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* May 37(1):67-75

Sears, R. C. (2004) The life cycle of C-myc: from synthesis to degradation. *Cell Cycle* Sep;3(9):1133-1137.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. and Kahn, D. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform.* Sep 3:246-251

Service, R. (2005) Structural biology. Structural genomics, round 2. *Science* Mar 11;307(5715)1554-1558

Shaanan, B. (1983) Structure of human oxyhaemoglobin at 2.1 A resolution. *J Mol Biol.* 171:31

Shindyalov, I. N. and Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* Sep 11(9):739-747

Sillitoe, I., Dibley, M., Bray, J., Addou, S. and Orengo, C. (2005) Assessing strategies for improved superfamily recognition. *Protein Sci.* Jul;14(7):1800-1810

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. and Hood, L. E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* Jun12-18;321(6071):674-679

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. and Hood, L. E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321(6071):674-9

Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J Mol Biol.* Mar 25;147(1):195-197

Smith, T. F. and Zhang, X. (1997) The challenges of genome sequence annotation or the devil is in the details. *Nature Biotechnology* Nov 15(12):1222-1223

Soding, J. and Lupas, A. N. (2003) More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* 25:837-46

Sowdhamini, R., Burke, D. F., Huang, J. F., Mizuguchi, K., Nagarajaram, H. A., Srinivasan, N., Steward, R. E. and Blundell, T. L. (1998) CAMPASS: a database of structurally aligned protein superfamilies. *Structure* Sep 15;6(9):1087-1094

Srinivas, H., Juroske, D. M., Kalyankrishna, S., Cody, D. D., Price, R. E., Xu, X. C., Narayanan, R., Weigel, N. L. and Kurie, J. M. (2005) c-Jun N-terminal kinase contributes to aberrant retinoid signalling in lung cancer cells by phosphorylating and inducing proteasomal degradation of retinoic acid receptor alpha. *Mol Cell Biol.* Feb;25(3):1054-1069.

Steigemann, W. and Weber, E. (1979) Structure of erythrocruorin in different ligand states refined at 1.4 A resolution. *J Mol Biol.* 127:309

Swindells, M. B. (1995) A procedure for the sutomatic determination of hydrophobic cores in protein structures. *Protein Sci.* 4(1):93-102

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nickolskaya, A. N., Rao, b. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* Sep 11;4(1):41

Taylor, W. R. and Orengo, c. A. (1989) Protein structure alignment. *J Mol Biol.* Jul 5;208(1):1-22

Todd, A. E., Orengo, C. A. and Thornton, J. M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol.* Apr 6;307(4):1113-1143

Todd, A. E., Marsden, R. L., Thornton, J. M. and Orengo, C. A. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol.* May 20;348(5):1235-60

Tronrud, D. E. and Matthews, B. W. (1993) Refinement of the structure of a water-soluble antenna complex from green photosynthetic bacteria by incorporation of the chemically determined amino acid sequence. *Photosynthetic Reaction Center* 1:13

Uberbacher, E. C. and Mural, R. J. (1991) Locating protein-coding regions in human DNA sequences by a muliple sensor-neural network approach. *PNAS* Dec 15 88(24):11261-11265

Valencia, A. (2005) Automatic annotation of protein function. *Current Opinion in Structural Biology* 15(3):267-274

van Dongen, S. (2000) Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm

Venter, J. C., Smith, H. O. and Hood, L. (1996) A new strategy for genome sequencing. *Nature* May 30;381(6581):364-366

Viterbi, A. J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* April 13(2):260-267

Vitkup, D., Melamud, E., Moult, J. and Sander, C. (2001) Completeness in structural genomics. *Nature Struct Biol.* Jun;8(6):559-566

Vogel, C., Berzuini, C., Bashton, M., Gough, J. and Teichmann, S. A. (2004) Supra-domains: evolutionary units larger than single protein domains. *J Mol Biol.* Feb 20;336(3):809-823

Vogel, C., Teichmann, S. A. and Pereira-Leal, J. (2004) The relationship between domain duplication and recombination. *J Mol Biol.* 346:355-365

von Mering, C., Jensen, L. J., Snel, B., Hppoer, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A. and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 33(database issue):D433-D437

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. and Borl, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* May 471:399-403

Wang, Y., Geer, L. Y., Chappey, C., Kands, J. A. and Bryant, s. H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem Sci.* 25:300-302

Watson, J. D. and Crick, F. H. (1953) A structure for deoxyribose nucleic acid. *Nature* 171:737-738

Webb, E. C. (1992) Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Enzyme Nomenclature. Academic Press, New York

Weisstein, E. W. (1999) Correlation Coefficient. From MathWorld -- A Wolfram Web Resource.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Pontius, J. U., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T. A., Wagner, L. and Yaschenko, E. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* Jan 1;33:D39-45

Wieser, D., Kretschmann, E. and Apweiler, R. (2004) Filtering erroneous protein annotation. *Bioinformatics* 20(1):342-347

Wilkins, M. H. F., Stokes, A. R. and Wilson, H. R. (1953) Molecular structure of deoxypentose nucleic acids. *Nature* 171:738-740

Wolf, E. I., Grishin, N. V. and Koonin, E. V. (2000) Estimating the Number of Protein Folds and Families from Complete Genome Data. *J Mol Biol.* 299:897-905

Wu, C. and Nebert, D. W. (2004) Update on genome completion and annotation: Protein Information Resource. *Hum Genomics* Mar 1(3):229-233

Wu, C. H., Huang, H., Nikolskaya, A., Hu, Z. and Barker, W. C. (2004) The iProClass integrated database for protein functional analysis. *Comput Biol Chem.* Feb 28(1):87-96

Wuchty, S. and Almaas, E. (2005) Evolutionary cores of domain co-occurrence networks. *BMC Ecol Biol* Mar 23;5(1):24

Yona, G. and Levitt, M. (2000) Towards a complete map of the protein space based on a unified sequence and structure analysis of all known proteins. *Proc Int Conf Intell Syst Mol Biol* 8:395-406

Zhang, C. and DeLisi, C. (1998) Estimating the number of protein folds. *J Mol Biol* 284:1301-1305

Zhang, S. Q., Yang, W., Kontaridis, M. I., Bivona, T. G., Wen, G., Araki, T., Luo, J., Thompson, J. A., Schraven, B. L., Philips, M. R. and Neel, B. G. (2004) Shp2 regulates SRC family kinase activity and Ras/Erk activation by controlling Csk recruitment. *Mol Cell.* Feb 13;13(3):341-55.

Zomorodipour, A. and Andersson, S. G. E. (1999) Obligate intracellular parasites: Rickettsia prowazekii and Chlamydia trachomatis. *FEBS Letters* 452:11-15